ERROR HANDLING IN MULTIMODAL VOICE-ENABLED INTERFACES OF TOUR-GUIDE ROBOTS USING GRAPHICAL MODELS

THÈSE N^{O} 3581 (2006)

PRÉSENTÉE LE 17 JUILLET 2006

À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Laboratoire de l'IDIAP

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Plamen PRODANOV

M. Sc. in Communication & Security Technology & Systems, Technical University, Varna, Bulgarie
et de nationalité bulgare

acceptée sur proposition du jury:

Prof. A. Skrivervik, président du jury
Dr A. Drygajlo, directeur de thèse
Prof. W. Burgard, rapporteur
Prof. R. Moore, rapporteur
Prof. R. Siegwart, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Lausanne, EPFL
2006

Acknowledgements

This thesis is a result from a collaboration project between the Speech Processing and Biometrics Group at EPFL and the Autonomous Systems Laboratory. The thesis would never be possible without the valuable help and guidance of my thesis supervisor Dr. Andrzej Drygajlo. He cared for my work, me and the other Ph.D students in his group like a real father. Thank you, Andrzej, for your time and wise advices, for your patience and support during hard moments, for your fair and very human attitude.

I would also like to thank heartily to Prof. Siegwart, the director of the Autonomous systems Laboratory and one of my thesis jury members. Without his support for my experiments in the form of space, equipment, the robot RoboX as well as my experience during Expo.02, my work would lack a lot of convincing arguments. Thank you also for inviting me to the numerous social events organized by your lab.

I wish to convey my special thanks to the other members of my thesis jury as well: Prof. Moore and Prof. Burgard. Thank you all for the atmosphere and the valuable scientific discussion during my oral examination.

I want to acknowledge also the names of several people that I worked with, and that helped in refining my ideas, as well as in solving software and hardware implementation problems. These are my colleagues Anil Alexander, Jonas Richiardi and Krzysztof Kryszczuk from the Speech Processing and Biometrics Group, as well as Guy Ramel, Bjoern Jensen, Nicola Tomatis and Mathieu Meisser from the Autonomous Systems Laboratory. I would like to thank at the same time to the secretaries Marie-José Pellaud, Marianne Marion and Chantal Schneeberger for being always of timely help with my administrative problems, as well as when I was simply absent-minded.

Special thanks to all the people that participated in my experiments and to the big Italian group that made my social life in Lausanne really great. Many thanks Lorenzo and Ion for introducing me to the taste of prosciuto and jamon. It kept me productive until the very last moments of writing.

At the end, I want to thank by heart my parents and my family. Thank you Jack, thanks Margo, for encouraging me to chase my dreams. Many thanks to my sister Sevda, her husband Joro and my cousin Svilen. Thank you all for your warm long- and short-distance love, support and advice.

Last, and very important, I would like to convey my tender thanks to the woman of my heart. Thank you, Vanya, for being next to me in every possible circumstances, giving me love, support and understanding. Thank you for the unique moments together, and for keeping me stepping firmly on the ground and being reasonable above all. It is You and my Family, that I dedicate this thesis to. To the people that carry the spirit of my grand-mother in their hearts.

Abstract

Mobile service robots are going to play an increasing role in the society of humans. Voice-enabled interaction with service robots becomes very important, if such robots are to be deployed in real-world environments and accepted by the vast majority of potential human users. The research presented in this thesis addresses the problem of speech recognition integration in an interactive voice-enabled interface of a service robot, in particular a tour-guide robot.

The task of a tour-guide robot is to engage visitors to mass exhibitions (users) in dialogue providing the services it is designed for (e.g. exhibit presentations) within a limited time. In managing tour-guide dialogues, extracting the user goal (intention) for requesting a particular service at each dialogue state is the key issue. In mass exhibition conditions speech recognition errors are inevitable because of noisy speech and uncooperative users of robots with no prior experience in robotics. They can jeopardize the user goal identification. Wrongly identified user goals can lead to communication failures. Therefore, to reduce the risk of such failures, methods for detecting and compensating for communication failures in human-robot dialogue are needed. During the short-term interaction with visitors, the interpretation of the user goal at each dialogue state can be improved by combining speech recognition in the speech modality with information from other available robot modalities. The methods presented in this thesis exploit probabilistic models for fusing information from speech and auxiliary modalities of the robot for user goal identification and communication failure detection. To compensate for the detected communication failures we investigate multimodal methods for recovery from communication failures.

To model the process of modality fusion, taking into account the uncertainties in the information extracted from each input modality during human-robot interaction, we use the probabilistic framework of Bayesian networks. Bayesian networks are graphical models that represent a joint probability function over a set of random variables. They are used to model the dependencies among variables associated with the user goals, modality related events (e.g. the event of user presence that is inferred from the laser scanner modality of the robot), and observed modality features providing evidence in favor of these modality events. Bayesian networks are used to calculate posterior probabilities over the possible user goals at each dialogue state. These probabilities serve as a base in deciding if the user goal is valid, i.e. if it can be mapped into a tour-guide service (e.g. exhibit presentation) or is undefined - signaling a possible communication failure. The Bayesian network can be also used to elicit probabilities over the modality events revealing information about the possible cause for a communication failure.

Introducing new user goal aspects (e.g. new modality events and related features) that provide auxiliary information for detecting communication failures makes the design process cumbersome, calling for a systematic approach in the Bayesian network modelling. Generally, introducing new variables for user goal identification in the Bayesian networks can lead to complex and computationally expensive models. In order to make the design process more systematic and modular, we

adapt principles from the theory of grounding in human communication. When people communicate, they resolve understanding problems in a collaborative joint effort of providing evidence of common shared knowledge (grounding). We use Bayesian network topologies, tailored to limited computational resources, to model a state-based grounding model fusing information from three different input modalities (laser, video and speech) to infer possible grounding states. These grounding states are associated with modality events showing if the user is present in range for communication, if the user is attending to the interaction, whether the speech modality is reliable, and if the user goal is valid. The state-based grounding model is used to compute probabilities that intermediary grounding states have been reached. This serves as a base for detecting if the the user has reached the final grounding state, or wether a repair dialogue sequence is needed. In the case of a repair dialogue sequence, the tour-guide robot can exploit the multiple available modalities along with speech. For example, if the user has failed to reach the grounding state related to her/his presence in range for communication, the robot can use its move modality to search and attract the attention of the visitors. In the case when speech recognition is detected to be unreliable, the robot can offer the alternative use of the buttons modality in the repair sequence.

Given the probability of each grounding state, and the dialogue sequence that can be executed in the next dialogue state, a tour-guide robot has different preferences on the possible dialogue continuation. If the possible dialogue sequences at each dialogue state are defined as actions, the introduced principle of maximum expected utility (MEU) provides an explicit way of action selection, based on the action utility, given the evidence about the user goal at each dialogue state. Decision networks, constructed as graphical models based on Bayesian networks are proposed to perform MEU-based decisions, incorporating the utility of the actions to be chosen at each dialogue state by the tour-guide robot. These action utilities are defined taking into account the tour-guide task requirements.

The proposed graphical models for user goal identification and dialogue error handling in human-robot dialogue are evaluated in experiments with multimodal data. These data were collected during the operation of the tour-guide robot RoboX at the Autonomous System Lab of EPFL and at the Swiss National Exhibition in 2002 (Expo.02). The evaluation experiments use component and system level metrics for technical (objective) and user-based (subjective) evaluation. On the component level, the technical evaluation is done by calculating accuracies, as objective measures of the performance of the grounding model, and the resulting performance of the user goal identification in dialogue. The benefit of the proposed error handling framework is demonstrated comparing the accuracy of a baseline interactive system, employing only speech recognition for user goal identification, and a system equipped with multimodal grounding models for error handling.

Keywords: Human-robot interaction, mobile tour-guide robots, voice enabled interfaces, multimodal fusion, dialogue error handling, multimodal grounding, Bayesian and decision networks

Version Abrégée

Les robots de service mobiles seront amenés à jouer un rôle de plus en plus important pour la société dans le future. Si de tels robots doivent être déployés dans des environnements réels et acceptés par la majorité des utilisateurs humains potentiels, l'interaction vocale devient très important. La recherche présentée dans cette thèse a trait au problème de l'intégration de la reconnaissance de la parole dans l'interface vocale interactive d'un robot de service, et en particulier d'un guide robotique.

La tâche d'un robot guide est d'engager un dialogue avec des visiteurs dans des grandes expositions, pour fournir le service approprié (par exemple montrer des parties de l'exposition) en un temps limité. Le point crucial lors de la gestion des dialogues du guide est l'extraction du but de l'utilisateur, autrement dit son intention de demander un service particulier dans chaque état du dialogue. Dans des conditions de grande exposition, les erreurs de reconnaissance vocale causées par le bruit ambient et l'attitude non-coopérative des utilisateurs n'ayant pas d'expérience préablable en robotique sont inévitables. Elles peuvent rendre difficile l'identification du but de l'utilisateur, ce qui peut mener à des échecs de communication. Pour réduire le risque de tels échecs, des méthodes pour détecter et compenser les échecs de communications dans les dialogues homme-robot sont nécéssaires. Durant les intéractions à court terme avec les visiteurs, l'interprétation du but de l'utilisateur à chaque état du dialogue peut être amélioré en combinant la reconnaissance vocale de la modalité parole avec de l'information d'autres modalités du robot. Les méthodes présentées dans cette thèse utilisent des modèles probabilistes pour effectuer la fusion de l'information provenant de la parole et de modalités auxilaires du robot afin d'identifier le but de l'utilisateur et de détecter les échecs de communication. Nous étudions l'utilisation de méthodes multimodales pour la compensation des échecs de communications détectés.

Nous utilisons le cadre probabiliste des réseaux bayésiens pour modéliser le processus de fusion de modalités tout en prenant en compte les incertitudes quant à l'information extraite de chaque modalité d'entrée durant l'interaction homme-robot. Les réseaux bayésiens sont des modèles graphiques qui représentent un fonction de probabilité jointe sur un ensemble de variables aléatoires. Ils sont utilisés pour modéliser les dépendences entre les variables associées avec les buts de l'utilisateur, les événements reliés aux modalités (par exemple la présence d'un utilisateur inférée depuis le scanner laser), et les paramètres observé des modalités fournissant des indices en faveur de ces événements. Les réseaux de Bayes sont utilisés pour calculer des probabilités *a posteriori* sur les buts d'utilisateur possibles dans chaque état du dialogue. Ces probabilités servent comme base pour décider si le but de l'utilisateur est valide (c'est à dire qu'on peut trouver une correspondance avec un service de guide, par exemple montrer la partie suivante de l'exposition), ou non défini, ce qui signale un possible échec de communication. Le réseau bayésien peut aussi être utilisé pour obtenir des probabilités reliés aux événements de modalité, révélant ainsi les causes possibles pour les échecs de communication.

L'introduction de nouvelles fonctionnalités relatives à l'utilisateur (de nouvelles modalités et les

attributions s'y rattachant par exemple) apporte des informations auxiliaires, permettant ainsi de palier à des échecs de communication. Le système résultant invoque, de par sa complexité, une approche systématique reposant sur une modélisation par réseaux bayésiens. L'introduction dans de tels réseaux de nouvelles variables visant à permettre l'identification de l'utilisateur conduit généralement à des modèles complexes et induit par conséquent des coûts de calculs élevés. Afin de définir un processus tout à la fois plus systématique et modulable, nous avons adapté des principes issus de la théorie des rudiments en communication humaine. Lorsque des personnes communiquent, elles résolvent des problèmes éventuels de compréhension par des efforts joints visant à atteindre un socle de connaissances communes (rudiments). Nous utilisons des topologies de type réseaux bayésiens, adaptés à des ressources de calcul limitées, pour définir un modèle d'états rudimentaire fusionnant les informations issues de trois modalités d'entrée (laser, vidéo et parole), afin de déduire des états rudimentaires possibles. Ces derniers sont associés à des évènements modaux évaluant si l'utilisateur est à portée de communication, s'il prend part à l'interaction, si la modalité de parole est fiable, et si la requête de l'utilisateur est valide. Le modèle d'état rudimentaire est utilisé afin de calculer les probabilités d'avoir atteint des états rudimentaires intermédiaires, ceci afin de déterminer si l'utilisateur a atteint l'état rudimentaire final ou si une séquence de dialogue réparatif est nécessaire. Dans ce cas, le robot guide peut exploiter les multiples modalités disponibles allant de pair avec la parole. Par exemple, si l'utilisateur n'est pas parvenu à atteindre l'état rudimentaire correspondant à une mise à portée de communication, le robot peut utiliser sa capacité de déplacement pour chercher à attirer l'attention du visiteur. Dans le cas où la reconnaissance de parole est considérée comme non fiable, le robot peut proposer comme alternative l'utilisation de l'interface tactile dans la séquence de réparation.

Étant données la probabilité de chaque état rudimentaire et la séquence de dialogue pouvant être exécutée dans le prochain état de dialogue, un robot guide dispose de différentes possibilités quant à la continuation possible du dialogue.

Si les séquences de dialogues possibles en chaque état sont définies comme des actions, il est possible d'introduire, comme moyen explicite de sélection d'action, le principe d'utilité maximum espérée (MEU), reposant sur l'utilité de l'action en fonction du but de l'utilisateur. Afin de prendre des décisions basées sur le MEU, des réseaux de décision sont proposés. Ces réseaux sont construits comme des modèles graphiques basés sur de réseaux de Bayes, et permettent d'incorporer l'utilité des actions à choisir par le robot guide à chaque état de dialogue. Ces utilités d'action sont définies compte tenues des tâches requises par la visite guidée.

Les modèles graphiques proposés pour la tâche d'identification du buts de l'utilisateur et la gestion des erreurs de dialogue lors de dialogues homme-robot sont évalués à partir d'expériences sur des données multimodales. Ces données ont été collectées durant le fonctionnement du robot guide RoboX au Laboratoire des Systèmes Autonomes de l'EPFL ainsi qu'à l'Exposition Nationale Suisse de 2002 (Expo.02). Les expériences évaluatives reposent sur l'utilisation de métriques au niveau des composants et des systèmes et sont constituées tout à la fois d'évaluations techniques (objectives) et faisant appel à l'utilisateur (subjectives). Au niveau des composants, la technique d'évaluation repose sur le calcul des précisions comme mesures objectives de la performance du modèle rudimentaire, ainsi que sur les performances lors de la tâche d'identification de l'utilisateur, à chaque état du dialogue. L'avantage apporté par le gestionnaire d'erreur proposé est démontré à travers une comparaison entre la précision d'un système de base n'utilisant qu'un système de reconnaissance pour la tâche d'identification de l'utilisateur, et celle d'un système équipé de l'un des modèles rudimentaires multimodaux permettant la gestion des erreurs.

Mots-clés: Interaction de humme-robot, robots-guide mobiles, interface vocale, fusion multimodale, gestion des erreurs de dialogue, rudiments multimodaux, réseaux bayésiens et de décision.

Contents

List of Figures

List of Tables

1 Introduction

A quote on understanding:

"Men don't want any new worlds. Only a mirror to see their own in. Man needs man!"
Stanislav Lem (1921-2006)

1.1 Mobile tour-guide service robots

Mobile service robots are physical agents that are designed to act in the real world, using their mobility to perform tasks useful for humans. Service robots perform some fixed number of services specific to the particular service robot application. The applications can include: medical, health care and rehabilitation robotics; commercial cleaning and household tasks; fast food service; aiding the handicapped and the elderly; entertainment, tour-guiding and educational robotics, etc. (Balaguer et al., 2004).

The focus in this thesis is on the tour-guide service robots. The tour-guiding services are related to presenting exhibits to visitors, while providing a tour in museums, mass exhibitions, trade fairs, etc. (Burgard et al., 1999; Drygajlo et al., 2003) Tour-guide robots need to interact with their users to decide on which exhibit to present. For this purpose the tour-guide robots are equipped with different modalities for user input (e.g. speech, interactive buttons, etc.) and output to the user, i.e. the modalities dedicated to exhibits presentation (speech, video, expressive face, etc.). The interaction with a tour-guide robot is of a short-term type. Visitors to mass-exhibitions are usually unprepared ordinary people, i.e. people without any prior experience with robotics. Intuitiveness and usability become very important when designing a communication interface for people that are not expected to have some prior experience with robots. Speech is an intuitive communication means for humans and for that reason development of system, which enable spoken interaction with tour-guide robot is very important research topic.

Spoken interaction with robots requires speech as input/output modality in the robot's voice-enabled interface. Speech recognition technology has gained performance in recent years that enables real-world applications (Huang et al., 2001). However, the state-of-the-art speech recognition techniques yield to recognition errors in noisy environment (Josifovski, 2002). Speech recognition errors

can lead to subsequent exhibit presentations that may not meet the user expectations and interest. In the case of robot's behaviors that do not correspond to the user expectations, the user can stop interacting and move away at any time.

This dissertation is focused on the problems of error handling when enabling speech-based interaction with a mobile tour-guide robot. In the introductory part, we provide background details on voice-enabled communication with tour-guide robots. We outline issues related to speech recognition that can lead to communication failures in the context of the tour-guide application. We then motivate the need of multimodal error handling techniques dedicated to reducing the risk of communication failures.

1.2 Voice-enabled communication with tour-guide robots

Voice-enabled communication between users and service robots is performed in the form of spoken dialogue. The participants in this dialogue are the robot and its user. The user is the person staying usually closest to the robot's speech input (microphone), communicating verbally with the robot. In the case of tour-guide robots users are visitors to the exhibition. There can be more than one person around the robot, however the robot is assumed to communicate with only one user at a given user turn in dialogue. Typically, the input modalities' sensors dedicated to interaction (e.g. microphone, video camera) are arranged to mimic anthropomorphic features, such as eyes, mouth or some sort of mechanic face facilitating interaction (Jensen et al., 2005).

We define the robot input modality as one of the possible inputs through which the robot can perceive and extract information from the external environment using sensors (Gibbon et al., 2000). The different robot sensors, such as microphone, video camera, laser scanner, etc. are related to different input modalities. The speech sensor of the robot can consist of a single microphone or of an array of microphones. In the case of sensors of the same type like the microphone array, we collectively refer to all of them as the same input modality sensor. An output modality can be defined as a functionality of the robot, which allows the robot to actively affect the state of the external environment. In the context of human-robot interaction the robot will affect the external environment with the goal of conveying information to people. For example, the speech output modality allows the robot to send acoustic message in the surrounding environment that can be perceived by a human user. The move modality allows the robot to move and change its position in the environment. The movement activity can be perceived by people and can attract their attention towards the moving robot.

Human-robot dialogues are constructed as a sequence of dialogue states (Gibbon et al., 1997) containing pairs of adjacent turns - one provided by each participant in the interaction (robot and its user). At each dialogue state the user has its turn in providing spoken input. Based on that, the robot has to infer the user goal, i.e. the user intention of requesting a given service. Exhibit presentations are the services provided by a tour-guide robot. According to the user goal, the tour-guide robot has to decide what service to perform in the next state. The number of possible services and user goals is assumed fixed by the particular service robot application. Typically, in the case of tour-guide robot the possible exhibit presentations and related user goals depend on the exhibition setting.

1.2.1 Limitations of speech recognition under noisy acoustic conditions

Voice-enabled interaction relies on speech recognition as the main source of information in the process of user goal identification. In noisy exhibition rooms, with many visitors that talk with the

robot and between themselves, the speech input modality may fail to provide reliable information for a particular user goal. Additionally, speech information may not be sufficient to reveal user presence in a room with groups of speaking people. The above limitations still prevent a widespread deployment of voice-enabled interface for the interactive systems of tour-guide robots.

When designing voice-enabled interfaces for tour-guide robots, we have to be aware that misinterpretations about the user goal can occur even in conversations between humans that should have almost perfect speech recognition abilities. Moreover, in the case of the tour-guide robot, the interaction takes place in spaces, where other speaking people than the user and the robot equipment itself can contribute to high levels of noise in the acoustic space. The speech input modality can deliver speech originating from a user, but also from other people speaking (passers by), causing errors in speech recognition.

1.2.2 Communication failures in spoken dialogues with tour-guide robots

Speech recognition errors can lead to incorrectly assigned user goals leading to subsequent exhibit presentations that may not meet the user expectations and interest. For example, a tour-guide robot, who continuously fails to recognize correctly the spoken user input will very probably drive its audience away. Moreover, literature on human-robot interaction has pointed out cases in which visitors try to confuse the robot for fun, e.g. blocking his way or using ambiguous answers to its verbal questions (Willeke et al., 2001; Drygajlo et al., 2003). Such user behaviors can additionally contribute to ambiguity and errors when the robot has to interpret the user goal using only speech recognition. Typically the recognition result consists of recognized words from the predefined keyword vocabulary. They are mapped to user goals corresponding to services offered by the robot (exhibit presentations). Incorrect speech recognition result can lead to user goals mapped with services that were not really requested by the user. The resulting dialogue continuation is in no way expected by the user and is the result of a communication failure. To avoid communication failures, tour-guide robots managing spoken dialogue with people should employ methods for speech recognition error handling, designed specially to face the needs of human-robot interaction systems.

1.3 Objectives of the thesis

The main objective of this thesis is to develop the error handling methods for reducing the risk of communication failures, when speech recognition systems are integrated in the human-robot interface of a multimodal robotic system. Since a mobile robot has to handle many other complex tasks along with interaction (e.g. localization, map-building, navigation planning, obstacle avoidance, etc.), the robot platform employs other input sensory modalities, such as different range finders and video cameras for tasks related to safe navigation. These modalities can provide complementary information that can be used along with speech in the spoken interaction. The proposed techniques have to be also tailored to limited computational resources of the on-board computer(s) and real-time operation.

This thesis states that in short-term interaction with visitors under adverse audio conditions, a combined interpretation of speech recognition with information from other available modalities can improve the user goal identification at each dialogue state. The central idea in the thesis is the statistical fusion of information from speech and other available robot modalities in order to prevent and compensate for the effect of the recognition errors in spoken interaction with tour-guide robots. The compensation is done by dialogue repair sequences that take advantage of the multiple robot modalities.

1.3.1 Statistical methods for fusion of modalities

At each dialogue state, during the interaction with a tour-guide robot, each exhibit presentation can be requested using particular spoken keywords. The recognized keywords are associated with corresponding user goals. Using only speech recognition result can lead to unreliable user goal assignment, due to factors such as acoustic ambient noise or unpredictable user behavior. Therefore, a deterministic association between the recognition result and the user goal can result in incorrect user goal identification. In this case, a probabilistic representation over the user goals and the observed recognition result is more informative. Auxiliary information from other available input robot modality can appear in this representation along with the speech recognition result, influencing the probability of each user goal value. In this way, a given user goal can become less probable in the presence of auxiliary modality information, pointing out at errors in the recognition result.

Therefore, the first objective of an error handling framework for human-robot interaction is to provide probabilistic representation for combining speech with information from other robot input modalities with the aim of detecting when the speech recognition result can lead to incorrect user goal identification.

1.3.2 Dialogue repair sequences

The statistical fusion of speech recognition with other modality information results in a probability distribution over the possible user goals. We define valid user goals as goals that can be directly mapped into services offered by the robot. However, it can often happen that the user behavior does not imply a valid user goal. To account for the cases when the user input cannot be interpreted as a valid user goal, we include an undefined user goal at each dialogue state. The undefined user goal often results from communication failures, such as in the case when out of the robot vocabulary words are used by the user in answering to the robot or when the user abandons the conversation with the robot. Having a probabilistic representation over the valid and undefined user goals the error handling framework has to make decision on which user goal to accept. If the undefined user goal has been selected as the best candidate, a communication failure has occurred. In this case, the error handling framework has to provide an interactive repair mechanism attempting to resolve the situation and identify a valid user goal in communication with the user.

Thus the second objective of the human-robot error handling framework is to provide tools for designing interactive sequences for communication failure repair at each dialogue state. The interactive repair mechanism exploits different input and output modalities.

1.4 Graphical models for error handling in human-robot interactive systems

1.4.1 Statistical fusion and graphical models

To address the first objective of the error-handling framework, i.e. providing probabilistic representation for combining speech with information from other robot input modalities, we have to investigate methods for multimodal information fusion. Since, the multimodal information in the case of service robots refers to the information extracted from the robot input modalities, we refer to multimodal information fusion as input modality fusion.

Input modality fusion can be defined as a process of combining information from multiple input modalities for achieving improved accuracies or more precise user goal inferences than in the case of a single modality (Mandic et al., 2005). In the case of tour-guide robots, inferring the user goal

during interaction can be done more accurately fusing auxiliary information from other robot input modalities along with the speech modality.

Probabilistic methods for modality fusion become attractive as they can explicitly represent the uncertainty intrinsic to each input modality information. They also provide a uniform way for handling different modality information. In the probabilistic approach to fusion, instead of combining physically different sensor information, we combine probabilities associated with this information.

In the very general case, probabilistic multimodal fusion is concerned with the problem of computing the probability over some variable of interest (query variable), given an assignment of observed (evidential) variables, associated with different input modality information. The observed variables can be related to feature values, extracted from raw modality data as well as from higher-level information, such as intermediate classification results. For example, higher level information is contained in the speech recognition result with respect to the raw sequence of speech feature values. The query variables are associated to events of interest that describe the particular random process under study. For example, one such variable can be associated to the different user goals in the process of human-robot interaction.

Recently, Bayesian networks, also known as probabilistic graphical models, have emerged as a unifying graph-based theoretical framework in the field of statistical modelling. All of the statistical models used in the context of multimodal information fusion, can be seen as particular types of Bayesian networks (Murphy, 2002).

Bayesian networks are directed graphical models that encode a joint probability function over a set of random variables. The variables can be discrete or continuous, and can be associated to observed features, intermediate classifier results or final decision variables such as the the user goal variable. The links in the networks are directed and are interpreted as causal influences from one parent variable to its children. Causal relations can be used to describe the probabilistic dependencies between the variables. The network topology in that way encodes efficiently the parametrization of a joint probability function over the model's variables.

Bayesian networks have well-defined algorithms for inference (Jensen, 1996; Murphy, 2002), i.e. algorithms for calculating probabilities over the variable of interest (e.g. user goal) given the observed evidential variables. Bayesian networks can be used to model the fusion of multimodal information for user goal identification. They can be also used for deciding if to trigger an interactive sequences for communication failure repair at each dialogue state.

1.4.2 Graphical models for dialogue repair

The main causes for human-robot communication failures are the ambient acoustic noise and the unpredictable user behaviors. In order to reduce the influence of the above two factors during interaction, principles from cognitive theories on grounding in human conversations can be used. The term "grounding" relates to the sufficient level of established joint attention, shared beliefs and understanding between the participants in the interaction. Grounding during interaction is established through getting spoken feedback from the user and the robot's perception of the level of user attention and understanding in the conversation. Grounding can also take into account the external environment factors and in particular the ambient noise.

The level of grounding can be modelled by a predefined number of grounding states that signify an increasing level of mutual understanding. These states can be associated with human behaviors perceived by the robot through its input modalities. Therefore, the process of grounding in human-robot interaction can be modelled with a state-based model. The model can exploit the multiple modalities available in the service robot system to provide evidence for reaching grounding states.

Bayesian networks combining speech and non-speech modality information, during user goal identification, can be used to estimate the probability that each grounding state has been reached. These probabilities can serve as a base for detecting whether the user has reached the final grounding state, or whether the user is failing to reach a particular state. In the later case a repair action can give the robot a second chance to achieve final grounding state and avoid a communication failure.

1.4.3 Decision networks for dialogue repair

Given the probability of each grounding state, a decision can be made on whether the grounding state is reached and what should be the dialogue continuation in the next dialogue state. The main purpose of a tour-guide service robot is to provide exhibit information to visitors. With this requirement in mind, the robot may have different preferences on the possible next dialogue states.

If the possible dialogue sequences at each dialogue state are defined as actions, the principle of maximum expected utility (MEU) (Russell and Norvig, 2003) can provide an explicit way of action selection, based on the evidence about the user goal and the utility of its associated dialogue state. Following the MEU principle the overall robot behavior can be seen as motivated by a high level goal of choosing the states that will maximize the robot accumulated utility. In this thesis, the principle of maximum expected utility can be implemented using extension to Bayesian networks known as decision networks (Jensen, 1996).

1.4.4 Complexity of inference with Bayesian networks

Bayesian networks can be used to create arbitrary statistical models using the causality relations behind the network variables. It should be noted however, that different Bayesian network topologies may require different time (Chapter 7) to perform inference. In the general case, with unrestricted network topology, inference can become NP hard to perform. NP hard (Non-deterministic Polynomial-time hard) refers to a class of computational problems that require non-polynomial number of computations in the size of the initial input. The size of the input in the case of Bayesian networks depends on the number of variables and the range of values that the variables can have. In this thesis we will show that the NP-hardness of inference can be greatly relaxed, when following some constraints in the phase of network construction.

1.5 Major contributions

In the emerging field of human-robot interaction with multimodal voice-enabled interfaces, we contribute by:

♢ Development of a voice-enabled interface for the purpose of tour-guiding by mobile robots.

♢ Providing a unified approach for speech recognition error handling in the framework of probabilistic graphical models.

In the light of the main objective of this thesis, our main contribution is:

♢ The development of new methods for speech recognition integration in an interactive voice-enabled interface of a service robot, in particular a tour-guide robot. The methods exploit statistical fusion of different input modalities for detecting communication failures, as well as interactive methods for subsequent recovery from communication failures using speech and other robot modalities.

Specific contributions related to statistical fusion of modalities are:

♢ Formulation of input modality fusion for identification of user goals and error handling in a spoken interaction with a tour-guide robot using the framework of Bayesian networks.

♢ Introduction of Bayesian networks for error handling models, based on low-level interaction grounding between the service robot and its user.

♢ Study and evaluation of Bayesian network topologies for robot modalities fusion tailored to limited computational resources.

Specific contributions related to dialogue repair strategies are:

♢ Development of methods for error handling in human-robot interaction, using dialogue repair sequences. The approach exploits the principles of decision theory and the potential of different modalities of the tour-guide robot.

♢ Introduction of a systematic approach for design and execution of multimodal repair sequences in human-robot interaction, based on the theory of grounding in human conversations.

1.6 Organization of the thesis

The thesis is divided in two parts as follows:

Part I: State of the art

♢ Chapter 2 provides a review of the state of the art in the domain of autonomous tour-guide robots, with a special emphasis on the human-robot interface design.

♢ Chapter 3 presents the state of the art techniques for error handling in spoken dialogue with emphasis on techniques that can be applied for handling recognition errors and subsequent communication failure repair in human-robot interaction.

♢ Chapter 4 reviews elements from the theory of Bayesian and decision networks, providing details on methods for efficient inference in networks containing both discrete and continuous variables.

Part II: Graphical models for error handling in human-robot interaction

♢ Chapter 5 presents the design methodology and field testing of the preliminary voice-enabled interface for the tour-guide robot RoboX. The results from this study outline requirements for designing spoken dialogue systems for tour-guide robots.

♢ In Chapter 6 a Bayesian network framework is used for combining input modality information for recognition error handling in human-robot spoken dialogue under adverse acoustic conditions.

♢ Chapter 7 is dedicated to the the problem of the strategy for recognition error repair, using dialogue sequences, where multiple modalities are used. We utilize principles from decision theory and cognitive theory of grounding in human conversations and Bayesian networks to create a multimodal grounding model for human-robot spoken interaction.

♢ Chapter 8 is dedicated to the evaluation of the multimodal error handling model, based on grounding presented in Chapter 7.

♢ Finally in Chapter 9 the main conclusions of this dissertation are summarized along with future research perspectives.

Part I

State of the art

Voice-enabled interfaces for mobile tour-guide robots

2

This chapter provides a review of the state of the art in the domain of autonomous tour-guide robots, with a special emphasis on the human-robot interface design.

The chapter starts with definitions about mobile service robots and mobile tour-guide robot as a special case. The second section presents particular examples of interactive autonomous robotic systems that were designed with the goal of guiding people and providing information in laboratory as well as real-world exhibition environments.

User input interfaces for tour-guide robots are based predominantly on tactile sensors (buttons, touch screens). We review existing interface solutions, based on utilizing speech recognition as a complementary solution to tactile input, outlining key problems that have to be addressed, when introducing speech recognition in the voice-enabled interface.

Our observation is that the existing research on the use of speech recognition for interactive tour-guide robots is still in its infancy. The recent trend in the broader domain of voice-enabled service robots is to utilize speech recognition in combination with other modalities in a multimodal interaction with robots.

The chapter ends with a brief discussion on the possible communication failures that may arise, when using speech recognition in mass exhibition conditions, without any mechanism for recognition error handling. The main reasons for communication failures, i.e. the noise in the exhibition room and different uncooperative user behaviors are described, along with possible perspectives for overcoming their effect in the voice-enabled interface.

2.1 Tour-guide robots

This thesis focuses on a particular service-robot application, i.e. the tour-guide robot. In one of the most frequently cited articles on tour-guide robots, (Burgard et al., 1999) the authors state that the primary task of a tour-guide robot is to give interactive tours through an exhibition, and to provide multimodal (e.g. employing audio and visual media) explanations to the various exhibits along the way. In a related work (Thrun et al., 1999b) the main tour-guide tasks are further refined into tasks related to: approaching people; interacting with them by replaying pre-recorded messages and displaying texts and images on onboard displays; safe and reliable navigation in un-modified and populated environment.

In this thesis, we will define tour-guide robot as:

Definition 1 (Tour-guide robot) *Mobile service robot whose main service (task/goal) is to provide exhibit presentations to visitors in a limited time, in the form of guided tour, using speech and other robot modalities in conveying information to its user (the visitor).*

As presented in (Schulte et al., 1999) the type of interaction provided by a tour-guide robot is spontaneous and short-term. Such interaction is typically limited in time due to the following reasons. First, visitors to mass exhibition are unlikely to spend all their time in the exhibition with the robot, since their interest and behavior can vary from investigative to collaborative as reported in (Schulte et al., 1999; Willeke et al., 2001). Second, the continuous visitor flow prevents visitors from staying in the exhibition room for a very long time. Third, the goal of the tour-guide robot is to provide exhibit information to as many visitors as possible. This requirement also calls for a limited time that visitors can spent with tour-guide robots, on average 10-15 min ((Schulte et al., 1999; Drygajlo et al., 2003; Jensen et al., 2002b)). In these circumstances the user interface has to fulfil the needs of users who lack prior exposure to robotics on one side, and the needs of a robotic system that has to handle tasks related to interaction as well as safe navigation in unstructured environment crowded with people.

Voice-enabled interfaces provide a communication means between people and the robot in a form of spoken dialogue. Voice-enabled interfaces are well-suited to the needs of the tour-guide interaction with people, in that they do not require prior user familiarity with the communication interface. However, in the interactive setting of noisy mass exhibitions, natural spoken dialogue becomes a difficult goal.

To define voice interface design recommendations, in the following section we make a review of existing voice-enabled interface solutions for service robots, and tour-guide robots in particular.

2.2 User interfaces for mobile tour-guide robots

The first attempt for designing a verbal interface for a service robot can be related to the period of the 60-ties, when the robot SHAKEY (Figure 2.1 (a)) (Nilson, 1984; Rajkishore Prasad, 2004) was developed. The robot was capable to move wooden blocks according to verbal commands introduced via keyboard.

Probably, the first tour-guide robot was Polly (Horswill, 1992), developed to offer guided tours in an office environment (the MIT's AI Lab). The user interface was based on visual cues. Typically the user would indicate its will to go on a tour by waving their feet.

In a more recent work, the potential of domain-restricted natural language is investigated for controlling a mobile robot (Torrance, 1994) in an office environment. The user interface supported natural language discourse with people entered via keyboard. The dialogue let user to name places,

ask questions about the robot's plan during navigation, and give the robot short and long-term goals.

These initial studies were about robots that are not truly voice-enabled (lacking real speech recognition component), however they outline the potential of natural language for communicating user goals to a mobile robot. They also motivate a special type of dialogue that is suited to the requirements of voice-enabled service robots, i.e. the task-oriented dialogue.

2.2.1 Task-oriented dialogues

Task-oriented dialogues are domain specific, i.e. they are organized as a joint activity to achieve a common task (McTear, 2002) between the dialogue participants. In the case of human-computer interaction, these common tasks can be e.g. reserving travel tickets, hotel rooms, car rentals, etc. In the case of service robots, the common task is related to providing the user with the desired service, which in the case of tour-guide robots is related to exhibits presentations.

The mobile robot MAIA (Antoniol et al., 1993) could carry objects from one place to another obeying simple spoken command phrases. MAIA was operating in an office environment, and was controlled by voice from a remote workstation. The idea of remote robot control (e.g. via web interface) appears later in the context of remotely controlled tour-guide robots in real exhibition conditions (Thrun et al., 1999b).

Another robotic assistant ROMAN (Hanebeck et al., 1997) was developed for providing services related to health-care, object manipulation and cleaning. The robot was able to operate semi-autonomously in an office environment, relying on spoken commands from a user using a workstation to command the robot. Natural language was used as an interface for specifying the robot task. The dialogue management component was based on the frame-filling concept (McTear, 2002), where input information consistence and sufficiency checking was performed, before translating the user command into a sequence of navigation and manipulation routines for the robot.

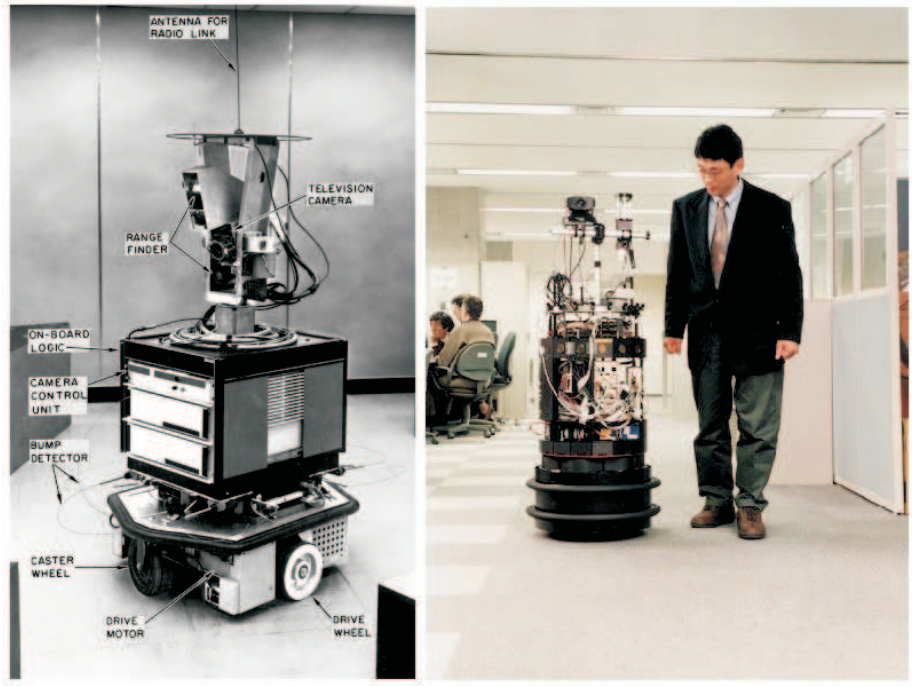
The above two examples demonstrate the use of task-oriented dialogues for spoken interaction with service robots. In the following section, we present a number of examples of tour-guide robots that were developed in the last 10 years. The focus of the review is on the user interface with a special emphasis on voice-enabled solutions and related problems.

2.2.2 Examples of tour-guide robots

Jijo-2 (Matsui et al., 1999) (Figure 2.1 (b)) was a tour-guide robot operating in an office environment that was fully voice-enabled and could offer peer-to-peer hands free communication with its user without the use of a remote workstation. The robot provided information about lab members, communicating with the user to acquire new information about its location. To ensure robustness of speech recognition against noise the robot was equipped with a microphone array. It also utilized ad-hoc techniques for recognition error handling based on confirmation of every spoken input provided by the user.

The robots described above were operating in laboratory conditions, communicating with people that were more or less familiar with the robot. The laboratory setting excludes presence of crowds of ordinary people. The particular application (e.g. fetch and carry tasks) itself required some initial familiarity with the robot and implied more long-term interaction type.

The first tour-guide robot deployed in real exhibition conditions was Rhino (Burgard et al., 1999) (Figure 2.2 (a)). Its user interface included speech synthesis for voice output and buttons for user input. It interacted with visitors of the Deutches Museum Bonn for 6 days. Although it did not feature real spoken dialogue, the experience during the operation of Rhino was valuable in that it



(a) (b)

Figure 2.1: (a) SHAKEY, (b) Jijo-2

stressed on the importance of human-robot interaction in real museum conditions. It became clear that navigational skills have to be combined with means for human-robot interaction in order to keep the visitor interested in the tour-guide robot presentations. The study outlined two main tasks of a tour-guide robot, i.e. attracting and keeping people involved on one hand, and using interaction to facilitate navigation through crowded spaces on the other hand. It also provided evidence that voice interfaces featuring speech recognition can be potentially beneficial, however existing solution has to be tailored to the noisy conditions and the short-term interaction style of tour-guiding in order to efficiently complement existing solutions based on buttons.

In a follow-up work, another tour-guide robot Minerva (Thrun et al., 1999b), (Figure 2.2 (b)) with improved interaction skills was deployed for 2 weeks period in the Smithsonian's National Museum of History (Washington, DC). The robot did not possess speech recognition capabilities, yet it was efficient in attracting people and driving away passers blocking the robot's path, utilizing a simple state-based mechanism for expressing emotions. The authors found this more basic level of human-robot interaction as efficient in contributing to the credibility of the robot character in the short-term interactive setting in the museum. However, according to the study (Schulte et al., 1999) people perceived the robot as being more similar in intelligence level to a dog than to a human tour-guide.

In a series of related studies (Bourgard et al., 2002; Matia et al., 2002; Alami, 2002; Maeyama et al., 2002) the idea of the web-operated mobile tour-guide robots is developed. In the TOURBOT and WebFair (Bourgard et al., 2002) projects three mobile robots are controlled by people via Internet. The robots also communicate with people on the exhibition site, using speech recognition of several phrases. In particular, the robot Albert that is similar in appearance to Minerva can recognize 20 phrases that are used to request information about the robot, the exhibition site, or

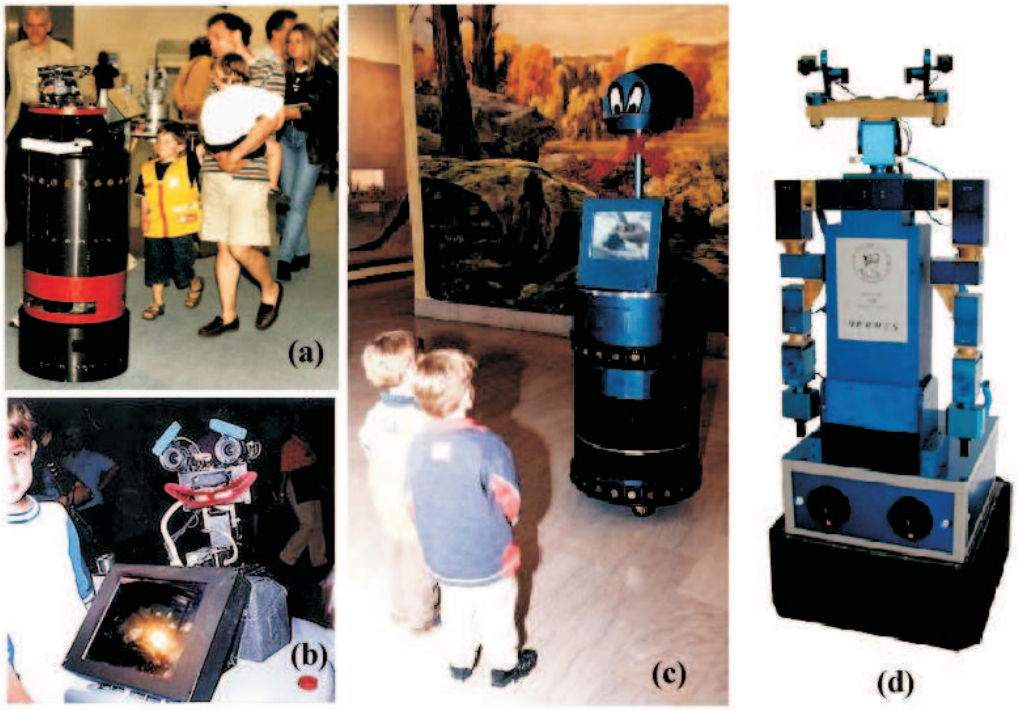


Figure 2.2: (a) Rhino, (b) Minerva, (c) SAGE/Chips, (d) Hermes

the time and the weather. The robot was using a commercial system for speech recognition and is reported to have achieved an overall recognition rate of 90%. Authors do not provide details on the particular dialogue structure, and how the evaluation was performed. The other studies describe interaction scenarios based on web user interface for remote robot control. In (Maeyama et al., 2002) the robot KAPROS is described as the remote eye for the user who is looking at an art exhibition via Internet. The robot operated mostly when the exhibition was closed to the public.

In (Vestli, 2002) the mobile tour-guide robot Museomobile is presented. The robot is capable of non-verbal interaction with visitors. The main goal of the designers has been a low-cost and reliable solution as far as navigation and maintenance are concerned. The navigation system is based on following a predefined path, using an inductive guiding stripe on the floor. The interaction with the visitor is achieved using special cubes that the robot can recognize. Depending on the cube attributes the robot has to present different exhibits. Exhibits are explicitly indicated by way-points on the robot path. The hardware architecture is based on hard-logic solution. Prerecorded audio files are used to communicate information to the user. There is additionally a possibility for changing these files when the robot is used for another exhibition. The authors report that the robot was well-appreciated by the museum stuff, since it worked reliably (no major faults) during five exhibit installations. On the other side, the lack of anthropomorphic features and the simple interaction interface did not provide the user with increased expectations concerning the robot intelligence. People perceived the robot mainly as a technical tool and were satisfied with its performance.

By the time when Rhino and Minerva were deployed Bischoff et al. (Bischoff, 1999) have developed a humanoid mobile platform Hermes (Figure 2.2 (d)). Hermes is equipped with a manipulator for fetch and carry tasks, and has been used as a tour-guide robot as well. The robot is equipped with a user interface based on speech. The robot is typically operating in a predefined area in the exhibition, where people can use off-board microphone to instruct the robot in performing fetch

and carry tasks using predefined sentences in almost natural looking conversation. The authors outline the system modularity as very important for dependable robot design and operation. They also comment on the robot's huge size (2m height) that could be reduced in the future for better acceptability by people and for safety reasons. The use of an off-board microphone in the case of Hermes can be seen as potentially limiting its application in real mass exhibition settings.

An interesting example of a voice-enabled tour-guide robot, deployed in real exhibition conditions is the tour-guide robot Eldi (Brito et al., 2001). The interactive dialogue between Eldi and its user is actually directed by an interactive monitor. Eldi is designed to perform shows and demos in a special region in the exhibition organized as a check-board. The robot plays different games (e.g. chess, the 8 puzzle, etc.) with the user, where the user communicates its moves using speech commands. The whole dialogue is guided by an interactive monitor on which the game board is displayed. The microphone itself is mounted off board close to the user. The whole interactive setting restricts the user to a predefined region in the exhibition room, so the conditions of audio acquisition and user behavior can be controlled. In this way a natural-looking and entertaining voice-enabled communication is performed. Although the work presents a nice example of how environmental conditions can be controlled and the problem of robust speech recognition can be addressed, we have to mention that the robot does not play a role of an active partner in the conversation. The real tour-guide in this case is the voice-enabled interactive monitor, and Eldi does not really provide tours.

Another nonconventional form of interaction between a collection of mobile exhibition robots and visitors was demonstrated at the World Expo 2000 at Hanover (Arndt, 2002). In this interactive setting, robots were performing complex collective figures regulating the visitor flow by dynamically changing the free space in the exhibition site. The goal of this non-conventional means for human-robot interaction was also to appeal to the artistic sensation of people as well.

All the robots described so-far were deployed in real exhibition conditions for a limited period of time (typically several weeks). In contrast to these examples the work in (Nourbakhsh, 2002; Graph and Barth, 2002) describes a long-term multiple robot installation. In this installation (Nourbakhsh, 2002) three robots have operated for 5 years in the Carnegie Museum of National History on different floors. The first robot SAGE/CHIPS (Figure 2.2 (c)) was a tour-guide robot and educator in the Dinosaurs hall. He provided audiovisual information to visitors about the dinosaur bone collections. The second robot SWEETLIPS was designed after considering experience with CHIPS and served as a tour-guide robot in the Hall of North America Wildlife. Since this section of the museum had lowest visitor traffic the robot was specially designed to attract visitors. The third robot JOE operated in the Atrium of the Heinz History Center, where JOE provided information and tours to permanent exhibits. All the three robots had a user interface based on buttons (touch screen) input and audio-visual output. Given the significant period of operation of these three robots, our preliminary voice-interface design considerations take into account the findings related to human-robot interaction outlined in this study (Section 2.2.4).

Another multi-robot installation is described in (Graph and Barth, 2002). Here the robots are three again, however the interactive setting is different. The robots have operated in the same hall of the Communication Museum Berlin since March 25^{th} 2000. They are interacting with people and between themselves as well. They have a specially designed outlook and personalities fitted to their role in the museum. The first robot The Inciting (Figure 2.3 (a)) is attracting and greeting visitors to the exhibition. The second one The Instructive (Figure 2.3 (b)) provides tours in the museum, giving explanations about the exhibits. The third one The Twiddling (Figure 2.3 (c)) acts as a child playing with a big ball. The robots speak to people, but do not really use dedicated buttons or speech user input. They can detect the users' activity and react to it. For example The Instructive

Figure 2.3: (a) The Inciting, (b) The Instructive, (c) The Twiddling

would react when people block its way by asking them for free space. The child-like robot would start to cry, when people are taking its ball, and at the same time The Inciting would ask people to give the ball back. The three robots also greet themselves when they pass near each other. All the three robots from the Communications Museum Berlin are based on the robotic platform of Care-O-Bot (Graf et al., 2004), produced by the Fraunhofer Institute of Manufacturing Engineering and Automation. The Care-O-Bot platform was primarily designed for personal robot assistant that helps people with everyday tasks in their homes.

After long-term experience with these three tour-guide robots, the developers outline that human-like communication via voice input is crucial for the full acceptance of robots in exhibitions and as social partners at home (Graph and Barth, 2002). In the context of social robot partner Roy et al. (2000) have investigated spoken communication, using partially observable Markov decision process to manage the interaction between patients and the "nurse" robot Pearl. Pearl was able to contact a patient, reminding about appointments, accompanying the patient to the appointment, and offering information of interest to that patient, related to e.g. weather forecasts, TV programs, etc.

Possibly the largest multiple tour-guide robot deployment in one exhibition room (Figure 2.4) was accomplished in the robotics exhibition during the Swiss National Exhibition (Expo.02) (Jensen et al., 2002a). Ten fully autonomous tour-guide robot RoboX were operating in the exhibition room for 6 months communicating with visitors showing different exhibits to them. The interactive interface comprised buttons input, with speech output combined with expressive face movements, as well as a domotic system for remote control of audio-visual devices. Two of the robots were equipped with microphone arrays and a speech recognition system in addition to the buttons, resulting in a fully voice-enabled interface for communication with visitors. The recognition system of RoboX is described in detail in Chapter 5 of this thesis. The evolution of the voice-enabled interface of RoboX is the main topic of research in the thesis.

More recent examples of mobile robotic platforms, utilized for tour-guiding are described in

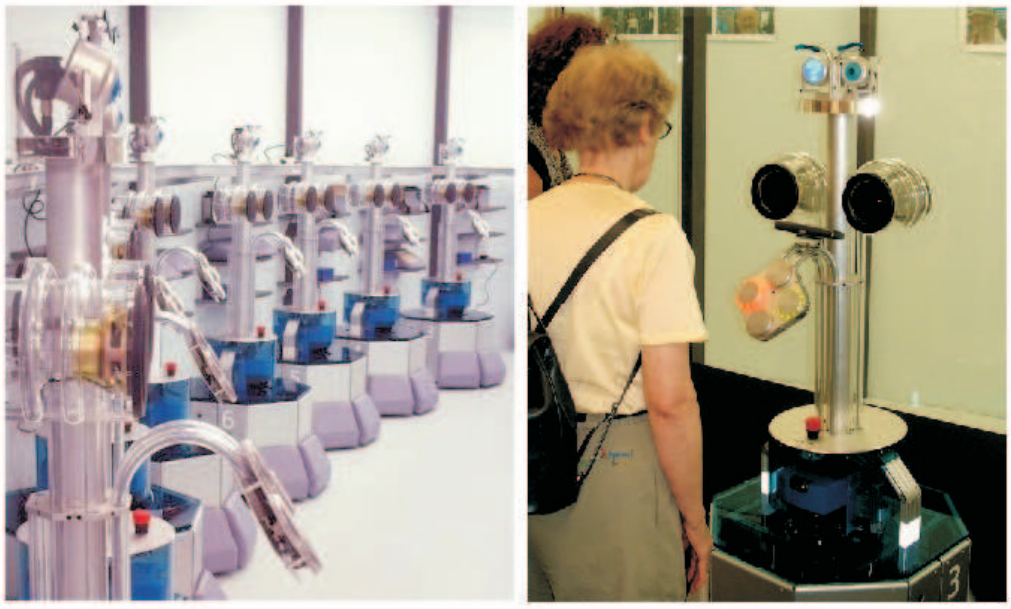


Figure 2.4: The tour-guide robots (RoboX) at Expo.02

(Haasch et al., 2004; Kulyukin et al., 2004; Clodic et al., 2005). Haasch et al. (Haasch et al., 2004) describe the mobile robot Biron (Figure 2.6 (a)) in an alternative tour-guide scenario in which the human user is showing the robot her/his home, familiarizing the robot with different objects in the home environment. The user interface is based on a multimodal approach for voice-enabled interaction, in which natural dialogue is combined with gesture recognition. For the purpose of speech recognition and sound-source detection Biron is equipped with a stereo microphone. The robot is able to detect its user using information from audio, video and laser sensors. Depending on the context of the conversation Biron can engage in gesture recognition in order to memorize a particular object that the user is pointing at. The dialogue model is based on finite state machines (FSM) that encode specific dialogues. The dialogue strategy is based on the frame filling method (McTear, 2002). Every different combination of the slots in a frame defines a state in the corresponding FSM. The task of the dialogue manager is to fill enough slots, given the user input, in order to meet the current dialogue goal (e.g. particular object specification from the home environment).

Kulyukin (Kulyukin et al., 2004) describes an application of a tour-guide robot for blind people aid. The robot RG (Robot Guide), presented in Figure 2.5, performs tours for blind people in indoor environments (office spaces). The user interface employs speech recognition for user input and speech synthesis for outputting information to its user. The output to the user combines verbal feedback as well as different audio signals. For example a bubbles sound indicates a presence of a vending machine offering beverage drinks. The important locations on the path of the tour-guide are indicated by passive tags detected from the on board active antenna. A special leash is used by the user in order to be lead by the robot during the tour. Speech recognition and synthesis is very useful in the case of blind people, as they cannot take advantage of visual cues, and the use of buttons or keyboard is not very convenient for them. However, the voice-enabled interface has to be very well designed to avoid recognition errors leading to wrong navigation task assignments. The author reports in his study on recognition errors, due to noisy speech recognition, that have resulted in RG suddenly deviating from a desired route. Similar problems appeared when the user

was engaged in a conversation by passers-by, and the continuously running speech recognition task has made the robot move without being commanded by its user.

Figure 2.5: The robot guide RG for blind people aid

Clodic et al. (Clodic et al., 2005) describe the tour-guide robot Rackham (Figure 2.6 (b)) deployed recently at the Mission Biospace exhibition in Toulouse, France. The exhibition features what could be an inhabited spaceship with a robotic guide inside. The user interface of the tour-guide robot Rackham is based on audio-visual output and user input through a touch screen. The robot's body consists of a mobile base and a mast with a helmet. The helmet is equipped with a pan-tilt camera, and another camera inside the helmet. The second camera is used for localization, while the pan-tilt camera is used for face detection. A ring of white LEDs are used to compensate for the difference in the illumination conditions throughout the exhibition, while the robot is performing face detection. An animated head serves as focal point for the interaction with the user. The face detection system is used by the robot to acknowledge user presence and to explain to the user how to operate the touch screen. The authors did not use speech recognition in their interface. However, they acknowledge the fact that visitors did not recognize that the robot was deaf, attributing to him speech recognition abilities as well.

The authors also outline the importance of the robot appearance in a social interactive setting such as tour-guiding. For example, human-like robots employing anthropomorphic features (e.g. an expressive face) tend to be more appealing, grabbing the attention of people. In this context humanoid robots could offer attractive solutions, however the visitors also tend to have raised demands related to the robot intelligence on facing more human-like robot.

In (Bennewitz et al., 2005) a humanoid tour-guide robot Alpha (Figure 2.7) is presented. The robot is using a multimodal interface for communicating with visitors, based on detecting sound sources and recognizing phrases from an audio channel, and detecting and tracking faces in the video channel. The interaction is performed using state-based dialogue and performing various robot behaviors to attract people and keep them involved in the interaction. The behaviors are executed by using a robotic face equipped with expressive eyes and mouth (Figure 2.7). The eyes

Figure 2.6: (a) Biron, (b) Rackham

can move and thus behaviors related to people eye tracking can be performed. In addition, the mouth and eyebrows can be used to express six basic emotions and a combination between them. In that way the robot is capable of human-like multimodal communication, where the robot emotions are regulated based on the presence or absence of visitors. At present only the robot head has been used in experiments with people during the RoboCup German Open 2005 in Padenborn. In order to use the robot as a real mobile tour-guide robot the head has to be mounted on a humanoid body. The authors plan to investigate the combined use of the head and the body gestures for more human-like multimodal interaction.

2.2.3 Speech recognition and the tour-guide robot task environments

From the presented examples of tour-guide robots, we can see that robots recently have employed different human-like anthropomorphic features in the interface design for more intuitive interaction with visitors. The initial robot constructions that were oriented to solve navigation problems would later feature a cartoon-like face (Thrun et al., 1999b; Jensen et al., 2002a) or an animated human face (Clodic et al., 2005), enabling the robot ability to express emotions intuitive to humans. The interface solutions also tend to evolve in the direction of multimodal interaction, acquiring multimodal input from multiple sensors and conveying information relying on multiple output modalities (audio, video, expressive face).

Researchers widely agree on the fact that the speech modality plays a crucial role in the tour-guide multimodal user interface, because speech can offer very convenient means for communication with humans (Kulyukin et al., 2004; Haasch et al., 2004; Graph and Barth, 2002; Drygajlo et al.,

Figure 2.7: The humanoid tour-guide robot Alpha

2003). However, speech recognition performance in noisy conditions is a bottleneck of the technology towards its widespread use for the needs of human-robot interaction with tour-guide robots (Kulyukin et al., 2004; Drygajlo et al., 2003). For this reason most of the tour-guide robots deployed in real exhibition conditions employ buttons or touch screens as input modalities in their user interfaces.

The literature on voice-enabled interfaces employing both speech recognition and synthesis is still rather sparse. This fact hints on a need for new design methodologies that explicitly address the problems of speech recognition integration in a mobile tour-guide robot platform.

2.2.4 Communication failures in tour-guide robot dialogues

Speech recognition errors in noisy exhibition rooms can result in communication failures in dialogue. The communication failures generally appear when the dialogue participants fail to understand each other. Participants to human-robot interaction may not understand each other because of the imperfect communication interface. For example, in the case of voice-enabled interfaces that are not adapted to noisy conditions, speech recognition errors may lead to a wrong interpretation of the service requested by the user. In this case, wrong services can be executed as presented in (Kulyukin et al., 2004). In this example, speech recognition plays an important role in enabling blind people to control a tour-guide robot. Speech recognition errors, however can lead to hazardous situation for visually impaired people, when the tour-guide robot changes route, obeying a wrongly recognized user command.

Communication failures may arise due to lack of sufficient user feedback as well. As explained in (Clodic et al., 2005) the use of a human-like voice communication with the help of the animated

face and speech synthesis can make the user believe that the robot can understand speech. Even when the robot can recognize speech, without a proper feedback it may be unclear for the user what she/he can say. Thus users lacking sufficient understanding of how the robot operates, and how the robot can be controlled using voice, may cause unreliable recognition using verbal expressions unknown to the robot.

Another source for communication failures in the tour-guide dialogue can be the typical visitor behavior as reported in the literature (Burgard et al., 1999; Thrun et al., 1999a; Willeke et al., 2001; Nourbakhsh, 2002; Clodic et al., 2005). People in exhibition conditions may not behave as instructed by the tour-guide robot. Visitors are generally described to be rather "destructive", when communicating with a tour-guide robot. For example, children try to establish interaction at more basic level becoming involuntarily "destructive", while pushing all possible buttons or climbing the robot platform (Drygajlo et al., 2003). Curious individuals may want to investigate how the robot operates in general, regarding the robot purely as a machine, and refusing to pay attention in the interaction. Most of the time people follow the robot, listening to the robot instructions. However, being aware of the fact that the robot is not a human, they may often quit interacting with the robot when some other people are calling them. For example, parents may follow their children, who have lost interest in the current tour. In these cases, the speech recognition will not provide the needed situation awareness, and the interpreted recognition result can result may produce weird robot behaviors, such as "talking to walls".

In order to reduce the risk of communication failure, when utilizing speech recognition in the voice-enabled interface of a tour-guide robot we need to exploit techniques for recognition error handling tailored to the needs of the tour-guide interaction setting.

2.3 Summary

In this chapter we have presented a series of tour-guide robot studies, focusing on the problem of interaction with visitors. The authors of these widely recognize the potential usability of speech recognition as an intuitive means of communication for people, however in typical noisy exhibition conditions with many not necessarily cooperative visitors, the use of speech recognition can lead to communication failures.

Given that in most of the studies the input user interface is based on tactile solutions, the review outlines the need for designing methodologies concerning voice-enabled interfaces for tour-guide robots, employing speech recognition as an input modality. The reviewed examples of tour-guide robots also outline that speech recognition has to be equipped with dedicated mechanism for speech recognition error repair, where multimodal spoken interaction may be of potential benefit.

Error handling methods in spoken dialogue with tour-guide robots

3

This chapter presents the state-of-the-art techniques in spoken dialogue error handling with emphasis on methods that can be applied for handling speech recognition errors and subsequent communication failure repair in tour-guide dialogue, i.e. a dialogue between a tour-guide robot and its user.

Section 3.1 is dedicated to methods for robust speech recognition in noisy conditions. Three main directions are presented in this context, i.e. techniques based on speech signal enhancement, techniques based on noise-robust speech parameterization, and model-based techniques for noise-robust recognition. These techniques can enhance the performance of the speech recognizer in noisy conditions. However, due to the partial information that the speech modality can provide about the user behavior in the tour-guide dialogue, the application of the above technique alone can fail to provide a reliable input for detecting user goals by the robot.

Section 3.2 in this chapter elaborates on existing dialogue-based techniques for speech recognition error detection and correction and their applicability in the the tour-guide dialogue settings. Many existing techniques for error handling in dialogue rely on ad-hoc solutions that may fail to provide the users of tour-guide robots with efficient communication means for conveying their user goals. In the case of tour-guiding, an efficient means of communication should not result in repetitive, time-consuming dialogue repairs, since such repairs are very likely to drive the user away in the short-term interactions in mass exhibition rooms. We then present existing systematic approaches for error handling in dialogue based on cognitive theories of grounding in conversation.

After motivating the need to investigate such systematic error handling techniques in the context of tour-guiding, we turn to reviewing existing error-handling techniques already applied in different human-robot interaction settings. We show that most of the existing methods for error handling in dialogue are solely speech-based and do not use available auxiliary information from other robot modalities. In this context, we outline the role of probabilistic methods for modality fusion that can be used to enhance the existing speech-based error handling methods.

3.1 Techniques for robust speech recognition in noisy conditions

Speech recognition performance in controlled acoustic environments has gained impressively low error rates recently (below 6 % for read speech in controlled conditions and vocabulary of about 20 000 words (Deng and Huang, 2004; Rajkishore Prasad, 2004; Wang et al., 2005)). However, the performance rapidly degrades in less controlled noisy conditions (Raj and Stern, 2005; Davis, 2002).

There are several reasons behind the drop in performance in non-controlled conditions of the current state-of-the art speech recognition technology that is based on statistical methods for speech modelling, i.e. Hidden Markov Models (HMMs). The main reason for performance degradation is the contamination with noise (additive, convolutional, reverberation) in real world acoustic environments such as populated exhibition rooms. Other reasons are related to speaking style, inter-speaker variations, as well as task-dependent reasons, such as typical behavior of visitors to mass exhibitions as already outlined in Chapter 2.

In this section, we concentrate on the effect of noise on speech recognition and the existing methods for achieving robustness of speech recognition in noisy conditions.

3.1.1 Types of noise

Additive noise typically results from the extraneous acoustic signal picked by the microphone along with the speech of the user. These can be sounds generated by the robot equipment. A tour-guide robot is equipped with motors and wheels as well as on-board computer(s) that produce substantial amount of noise. In addition, other people than the user, as well as other robots using synthesized speech in a multi-robot installation, contribute to high levels of noise that is similar to speech (babble or cocktail party noise).

The additive noise is additive to the useful speech signal in the time domain. It can be stationary, changing with time, as well as impulsive.

Convolutional noise is related to the transformation that speech undergoes while propagating through a given transmission channel, e.g. microphones, amplification equipment, etc. Convolutional noise is multiplicative to the speech signal in the spectral domain.

Reverberation noise results from the summation of the useful speech signal with its multiple reflections in the exhibition room at the point of speech capture. While the additive and convolutional noise can be assumed independent from the original speech signal, the reverberant speech is strongly related to the original speech signal.

The unpredictable manner in which noise in the environment can affect speech, contribute to significant mismatch in the statistical properties of the speech used for training of the recognition system and the speech to be recognized by the same system in operating conditions. Therefore, recognition systems that are used in the real-world applications have to be robust to the different effects of the different types of noise.

There are three main groups of methods, described in the literature, proposed to achieve the desired robustness of speech recognition system against noise, i.e methods related to speech enhancement, methods related to extracting speech features robust to noise, and methods related to changes applied to the speech recognition models.

3.1.2 Speech enhancement

Speech recognition systems do not use directly the time representation of the speech signal. They use feature vectors extracted from consecutive signal segments taken at equal time intervals. These

segments called windows have a fixed length, for which speech signal is assumed stationary (typically 25ms) (Huang et al., 2001). The signal segments can overlap and can be multiplied with a special window function (e.g. Hamming window). The speech features, extracted from these segments, aim at capturing pertinent and compressed information related to the vocal tract configuration.

Speech enhancement techniques try to reduce the difference in the statistical properties between the clean and noisy speech features using some $a-priori$ information about the properties of speech, noise or how they are combined. The primary goal of speech enhancement is speech de-noising. Therefore, these methods are not guaranteed to really improve the performance of a recognizer trained on clean speech that is supposed to operate on enhanced speech. For this reason, speech recognizers are trained after performing the particular enhancement over the clean speech used for training as well (Josifovski, 2002).

Speech enhancement is attractive, because it does not require any change in the speech recognition system. In the remainder of the section, we describe speech enhancement methods starting with those that can be already applied during the audio signal acquisition using microphone arrays. We then continue with the most common methods that can be applied as a preprocessing of the acquired speech signal.

Speech enhancement and audio signal capture

Noise cancellation can be already applied during the speech acquisition phase. This initial enhancement step is very useful as it typically comes "for free", i.e. without any computation required from the computer that is used to run the recognizer.

The input speech modality for spoken communication between visitors and mobile tour-guide robots requires on-board microphone sensors. Microphone arrays with their capability to perform precise spatial filtering (beam forming) of the acquired signal and de-reverberation become a very attractive choice for voice-enabled robots (Matsui et al., 1999; Choi et al., 2003).

The primary goals of microphone arrays are finding the position of a sound source, and improving the signal to noise ratio (SNR) of the captured audio signal (Huang et al., 2001). As the speaker is generally in some distance from the microphone, noise contaminates the captured speech. By playing with the distinct microphone spatial configuration we can achieve high microphone array directivity, performing additional spatial filtering of the background noise (Choi et al., 2003).

In this thesis we use a commercially available microphone array: Andrea DA-400 2.0. The microphone is equipped with the Andrea's Digital Super Directional Array 2.0 and PureAudio 2.0 noise-cancelling technologies. Optimized to filter out background noise and perform de-reverberation, the DA-400 comes embedded with a digital signal processor chip, along with four microphone chips. The specifications sheet for DA-400 2.0 is presented in Appendix A.

Methods based on signal preprocessing

The signal captured by the microphone can be additionally pre-processed for removing noise before performing speech recognition.

Spectral subtraction is the dominant method for removing noise from the speech signal. It relies on some average noise spectrum estimation. The estimated spectrum is then subtracted from the incoming speech signal short-term spectrum. The enhanced signal spectrum is used further in the feature extraction phase. The assumption is that the noise is stationary. Detailed descriptions of spectral subtraction can be found in (Berouti et al., 1979; Renevey, 2000; Josifovski, 2002).

Wiener filtering is commonly used as an alternative or complementary technique to spectral subtraction (Vaseghi and Milner, 1997) for removing additive noise.

A comprehensive list of other speech enchantment techniques is presented in (Renevey, 2000; Josifovski, 2002).

3.1.3 Robust features

Another set of methods do not aim at cleaning the speech features from noise, but extracting features already robust to noise. In other words extracting features that are not affected by noise.

The predominant feature type used in speech recognition is the mel frequency cepstral coefficients (MFCCs) (Figure 3.1). These features are derived after applying a Fourier transform based filterbank designed to give approximately equal resolution on a mel-scale. The mel-scale is inspired by properties of human auditory system. The emphasis is on a better sensitivity at lower frequencies to the expense of lower sensitivity to high frequencies.

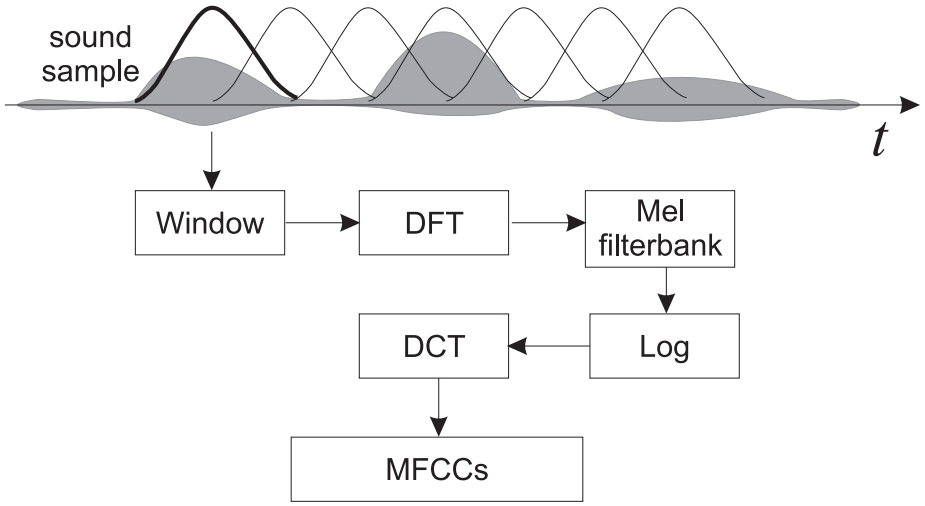


Figure 3.1: Block diagram of MFFC computation

The filters in the filterbank are triangular. The filtering results in multiplying each discrete Fourier (DT) magnitude with the filter gain. Afterwards all values per filter are added (accumulated). Typically before the accumulation, power values (squared magnitudes) can be computed. The number of triangular filters depend on the speech frequency band (Young et al., 2002). After applying the filter bank, we get filterbank coefficients. To get the log-filterbank coefficients, a logarithm is taken of the filterbank coefficients. This step also mimics human logarithmic perception to the increase in speech volume. Finally, to derive the MFCCs the log-filterbank coefficients are transformed into the cepstral domain, using discrete cosine transform. Typically the low-order 12 to 15 MFFCs are taken as final speech features, as they describe in a compressed form the spectral envelope of the short-term speech spectrum.

MFFCs are shown to perform better in noisy conditions that spectral representations (e.g. filterbank coefficients) (Josifovski, 2002). However, MFCCs also get contaminated by additive and convolutional noise. If the convolutional noise component is constant or slowly varying, it appears as a bias in the time evolution of the MFCCs values.

Cepstral mean normalization can be used with MFCCs to remove constant convolutional noise. Such noise may result from the transfer function of the microphone or the transmission channel through which speech is communicated. Applying *cepstral mean normalization* results in features robust to convolutional noise that does not change rapidly over time. In that way features can

become much less sensitive to different microphone equipment (Stern et al., 1997).

To combat more efficiently the additive noise component, different feature representations have been designed taking inspiration from the human auditory system. A special form of linear predictive analysis known as perceptual linear prediction (PLP) (Hermansky, 1990) is more effective in obtaining noise resistant features than the ordinary LP. The idea in PLP is to fit the poles to the warped mel-scale spectrum, rather than the linear one.

In addition, two other properties of human hearing are incorporated in the PLP feature extraction method. The critical band analysis is followed by equal loudness pre-emphasis and intensity-to-loudness conversion (by taking the cubic root of the critical filter bank values) before the LP coefficients are calculated (Hermansky, 1990). A discrete cosine transform is applied to the LP coefficients to result in the final PLP coefficients.

In order to address both robustness against additive and convolutional noise PLP features were used as a base for computing RASTA (RelAtive SpecTrA) features (Hermansky and Morgan, 1994). The idea of RASTA is to suppress any speech component that changes more slowly or more quickly than the "typical" range of change for the clean speech signal. RASTA is again motivated by human perceptual system. Humans tend to respond to changes in the intensity value of the input speech rather than the absolute value of the input speech.

In RASTA processing of speech the final spectral components derived from the filterbank are compressed and filtered, where the trajectory of the filterbank output over time is itself filtered. The filters are designed to suppress constant factors in each filterbank output. The last processing step is all-pole modelling like in PLP. The filtering described above operates in log-spectral domain. Filtering slow varying components in this domain corresponds to filtering convolutional noise in time domain. However, filtering in the log-spectral domain does not compensate for additive noise component. It has been demonstrated that high-pass filtering of the envelope of the bands can be effective for additive noise removal (Hirsch et al., 1991). This effect is achieved in J-RASTA by adding a small constant to the output of the filterbank before the log compression. This amounts to noise masking. J-RASTA-PLP features appear to be effective with wide range of noises addressing both additive and convolutional noise (Hermansky and Morgan, 1994).

Another technique, based on the use of dynamic features, is also shown to improve the robustness of speech against noisy conditions. *Derivatives* can remove slowly changing convolutional noise, when applied in the cepstral domain. The same is true with additive noise when applied in the spectral domain. The derivatives of the "static" features are computed either by simple difference or via regression (Furui, 1986). The use of first and second order derivatives in combination with the "static" MFCCs has been shown to improve robustness of the models trained on clean speech against speech produced in noisy conditions (Hanson and Applebaum, 1990).

3.1.4 Model-based techniques

One simple way to achieve robustness in recognizing noisy speech by manipulating the recognition models only, is to have access to the particular deployment conditions and to train the recognition models with noisy speech. However, this approach may not lead to acceptable results especially in very noisy conditions. The models trained on noisy data may lack sufficient discrimination and lead to unacceptable word error rates. In the literature, there are reported results on recognizers trained at several SNRs, showing an increase of the average word error rate from 1.48 % to 38.29 % on test speech contaminated with noise at 0 dB global SNR (Josifovski, 2002).

Another possibility for improving the recognition performance with noisy speech is to perform recognition models adaptation. There exist different methods for supervised and unsupervised model adaption with statistical models for speech recognition (e.g. HMMs) (Woodland et al., 1996a,b).

They typically use data-dependent linear transformations on the model parameters that results in better match in a statistical sense between the model and the observed noisy speech.

The above methods do not need explicit modification of the recognizer and its models. More complex models can achieve better performance by employing explicit statistical models for noise, speech and how they are combined. In particular, the use of parallel model combination (PMC) technique in the HMM framework has been reported to improve significantly the recognition performance in very low SNR conditions (Gales, 1995). An alternative approach inspired by the phenomenon of masking and the redundancy of speech spectral representation is the missing features approach to speech recognition (El-Maliki and Drygajlo, 1999; Renevey and Drygajlo, 2000; Josifovski, 2002). It is reported to have comparable performance to PMC (Renevey, 2000) at very low global SNR. Both PMC and missing feature approaches require explicit modification of the standard models and algorithms used for speech recognition.

In this thesis, we will use a combination of the approaches mentioned above to achieve noise robustness of speech recognition in mass exhibition conditions. In selecting the particular methods we are influenced mainly by requirements for implementation simplicity and low computational costs. Given the above requirements, we utilize a microphone array Andrea DA-400 2.0 to enhance the acquired speech against the effects of additive noise and room reverberation. In order to address the problem of convolutional noise, we chose to use MFCCs plus their first and second derivatives. For additional robustness against noise, in our recognition experiments (Chapter 5 and Appendix B) , we use also the methods of recognition model retraining with noisy speech and supervised model adaptation (Young et al., 2002).

3.2 Dialogue-based methods for handling speech recognition errors

The techniques for robust speech recognition can lead to better recognition performance with noisy speech. Speech recognition errors however, can still occur in noisy conditions. Even simple utterances (single words) pronounced in adverse (noisy) acoustic environment can be misclassified in the recognition process (Turunen and Hakulinen, 2001). It is often said that the major problem in voice-enabled human-computer interfaces is the interface inability to detect and correctly handle different speech recognition errors. Thus, error management in real-world spoken dialogue applications is crucial for successful human-robot interaction. However, most of the current tools for speech application development do not have decent support for error management and their performance depends only on speech recognition error rate in given conditions (Turunen and Hakulinen, 2001).

Speech recognition errors, if left without any error management, can contribute to two types of communication problems, i.e. non-understanding and misunderstanding (Skantze, 2003). In the case of spoken communication with service robots a non-understanding occurs, when the robot is unable to interpret the recognition result into any meaningful user goal (service request). A misunderstanding is related to incorrect user goal (service) assignment, due to recognition errors.

Error management is usually separated into the following phases: error prevention, error detection and error correction (Turunen, 2004). Dialogue-based methods of *error prevention* can be related to controlling the initiative in dialogue in order to reduce the risk for recognition errors. In this case, mixed initiative dialogue with more general-purpose recognition grammar can be switched to system-initiative dialogue with more restricted grammar (McTear, 2002). Errors can also be prevented, when offering context sensitive help to the user. Error prevention can take place in the stage of the recognition system design. As described in the previous section, different techniques for noise robust speech recognition can be employed, along with a careful design of the system vocabulary

words to minimize the chance of recognition errors. The above methods for error prevention can be directly applied in the case of human-robot interaction.

Error detection is related to the dialogue system ability to detect an error (recognition or understanding error, i.e. non or misunderstandings). Error detection is also related to detecting the user attempts to correct an earlier dialogue sytem error (e.g. wrong implicit confirmation). In many applications, error detection is left to the user, however the system should be also able to detect errors (Kamm, 1994).

Finally, *error correction* is related to the execution of different error repair sub-dialogues.

Luperfoy and Duff (1996) and later Turunen and Hakulinen (2001) have presented a finer classification of the phases of error handling, however in this thesis we will use only the three phases described above (prevention, detection and correction).

3.2.1 Detecting errors in spoken dialogue systems

Confidence measures for error detection

Detection of recognition errors typically relies on confidence measure from the speech recognition engine to asses the reliability of the current recognition result (Jiang, 2005; Torres et al., 2005; Cox and Dasmahapatra, 2002). Initially research in confidence measures concentrated on frame, phoneme and word level (Cox and Dasmahapatra, 2002). Confidence measures were built based on the recognition score (likelihoods), measures derived from the N-best hypothesis lists and different combinations derived after applying summation, product rule or neural networks (Garcia-Mateo et al., 1999). More recently, confidence measures began to utilize information from other dialogue system components than speech recognition. San-Segundo et al. (San-Segundo et al., 2000) used phonetic, language model as well as parsing features to detect misrecognized words and out-of-domain utterances, using neural network classifier. The study in (Walker et al., 2000) describes a system for miss-understanding detection based on decoder, dialogue management and system specific language-understanding features. In related studies (van den Bosch et al., 2001; Carpenter et al., 2001) authors present machine learning algorithms for error detection operating with features that can be found in most spoken dialogue systems at different levels (decoder, parser and dialogue management). In detecting misunderstandings, auxiliary features related to prosody can be helpful (Hirschberg et al., 2004). The general conclusion is that components other than speech recognizer, when available, can provide useful information for the confidence measure calculation. It has been shown however that confidence measures alone may not be enough for detecting all recognition errors (Krahmer et al., 2001; Sturm et al., 2001).

Using keyword spotting for error detection

In addition to confidence measures, detection of recognition errors resulting in non-understanding can be done using word spotting recognition techniques (Wilpon et al., 1990). The task of word spotting is to detect a set of keywords within the speech signal that can contain also other words, and in general acoustic phenomena other than the speech. These extraneous acoustic phenomena are modelled with the help of special models for out-of-vocabulary words, i.e. ”garbage models”. As a by-product in word spotting, non-understanding can be detected, when only ”garbage models” score as the most likely explanation of the current speech input.

Garbage models can be created using existing recognition models (Wilpon et al., 1990; Renevey et al., 1997) (Appendix B). With slightly modified speech recognition decoder, keyword spotting can be performed without the need of explicitly defined garbage models (Caminero et al., 1996; Silaghi and Bourlard, 1999; Silaghi, 2005). The advantage of the later technique is in the use of optimal

confidence-based methods for keyword classification. In the case of the explicit use of garbage models, filtering of non-keyword acoustic units is done in more ad-hoc manner, using e.g. penalty factor to tune the garbage grammar in testing conditions (Appendix B). The garbage models are typically constructed using existing or specially designed filter HMMs (Hidden Markov Models). In the case of specially designed HMMs, a speech corpora is needed to train the filter HMM. Word-spotting techniques with explicit garbage models may be less accurate than the methods without such models, but they work faster and do not need modification of the existing recognition search algorithms (Silaghi, 2005). In this way only grammar modifications may be needed with the existing recognition solution.

Multimodal error detection and prevention

The effect of recognition errors can also be reduced by using combination of speech and non-speech modalities. Oviatt (Oviatt et al., 2004; Oviatt, 1999) investigated speech-based multimodal interfaces and found that multiple modalities such as speech recognition and computer pen input can disambiguate each other. A multimodal interface can thus reduce the recognition errors significantly, for example in achieving similar recognition accuracy with non-native speakers as with native ones. Alternative modalities to speech that do not necessarily require keyboard, can enhance the usability of spoken dictation applications as reported in (Suhm et al., 2001).

Investigating the use of different input modality information is very convenient in the case of mobile tour-guide robots as such information is already available on the inherently multi-sensor robotic platforms.

3.2.2 Error correction strategies

Error correction strategies are used to correct the detected understanding errors (miss and non-understandings). The most common correction strategies employed in dialogue are explicit and implicit confirmations. Explicit confirmations require an extra turn in dialogue. Nevertheless, they are more reliable in their outcome than the implicit ones, because if implicit confirmation is inaccurate, detecting the following user correction attempts is much harder (Krahmer et al., 2001). Krahmer and Swert have investigated these two error correction strategies, identifying a set of positive and negative cues that people use in response to each of them.

In the short-term spoken interaction between visitors and tour-guide robots explicit confirmation would be much more preferable. The error-correction component in such interaction setting is usually a short dialogue, which can include a yes/no confirmation question or explicit request for repetition. In the error correction component additional modalities can be combined with speech for better intelligibility in the noisy exhibition room. If the error correction model is not well designed in noisy acoustic conditions, we can end up in repetitive error correction combinations. Signalling misunderstanding through these frequent error corrections can be very frustrating and can give the user an impression of a dialogue failure (Skantze, 2003). A frustrated user might leave the robot before completion of an error correction sequence. Hence, the ad-hoc use of dialogue error-correction techniques without considering the state of user attendance and the speech modality reliability can make a speech-based interaction with a service robot very inefficient.

Two main approaches exist, in general, for making decision regarding the possible error-correction strategy. The first approach is based on ad-hoc, empirically derived policies based on assessing values of different recognition confidence measures (Bohus, 2004). These techniques may fail to generalize well, and may need tuning for each new application. In contrast to them, theoretically inspired methods with their systematic approach to error correction may bring a greater benefit

when designing a voice-enabled interface for the tour-guide robots. In the following section we review the main systematic approach utilized in the domain of dialogue system and error handling modelling, i.e. the model of grounding in conversation.

3.3 Theory of grounding in conversation

The term grounding was used initially in the fields of psychology and cognitive science to explain the collaborative aspects of human-human interaction (Clark and Schaefer, 1989). In human-computer interaction the idea of grounding was later introduced to represent explicit, theoretically-motivated heuristics for the strategy of error handling in dialogue (Brennan and Hulteen, 1995).

3.3.1 Error handling models based on grounding

In the grounding theory the model of dialogue error handling is represented as an incremental process of establishing a common ground, i.e. level of understanding between the participants in the conversation (e.g. the user and the robot in our case). The common ground is related to the state of achieving sufficient understanding between the participants for the purpose of the conversation. In a collaborative dialogue setting, the state of sufficient understanding is closely related to the evidence that what is being said by the speaker is understood by the listener(s) considering the current purpose of the conversation. Such evidence is provided by the explicit and implicit feedback between the participants in the conversation. The feedback can be negative - signaling misunderstanding or positive - signaling increased level of understanding and finally agreement. Based on the needed evidence of mutual understanding, people may employ grounding actions. In human-computer interaction the dialogue error corrections can be seen as such actions (Brennan and Hulteen, 1995).

In their seminal work Clark and Schaefer (Clark and Schaefer, 1989) introduced a state model to represent the incremental process of grounding in a collaborative conversation between dialogue participants. In this model the level of sufficient understanding is explicitly represented by a set of states that an addressee R attributes to a speaker U and an utterance u. The state model is depicted in Table 3.1.

State	Description
State 0:	R did not notice that U uttered any u
State 1:	R noticed that U uttered u
State 2:	R correctly heard u
State 3:	R understood u

Table 3.1: Unimodal state model of grounding in conversation

All the states have to be reached in order to consider the current participant dialogue contribution as grounded. Whether a state has been reached depends on the evidence provided by the positive feedback from the speaker as well as by environmental factors related to the acoustic noise. The need for grounding actions arises whenever R has failed to reach one of the states in the model. In the case of a human-computer dialogue the speech modality should provide all the evidence for inferring the four grounding states in Table 3.1. Therefore we refer to this model as the unimodal grounding model. The unimodal grounding model was further extended by (Traum, 1999; Traum and Dillenbourg, 1998) who have proposed the conversational/grounding acts model, contributing to the taxonomy of speech acts with grounding-related acts. The authors proposed a quantitative model of the utility of a grounding acts, based on the value of a grounding criterion measure, the

added effect of the grounding act and its cost. Brennan and Hulteen (1995) have also extended the original grounding model with additional states and related grounding actions, commenting on the effect of the grounding criterion on selected grounding actions.

In summary, all the above studies concentrate on grounding using only speech as a communication medium. Their grounding models and definitions for the grounding criterion measure provide only specific solutions to the particular study.

3.3.2 Graphical models for grounding

Horvitz and Paek (Horvitz and Paek, 2001; Paek and Horvitz, 1999) have proposed a computational model for the process of unimodal grounding motivated initially by the Clark and Shaefer architecture. In this model they regard grounding and error handling in dialogue as a process of making decisions under uncertainty in a four-level architecture called the "Quartet". The uncertainty in taking a decision can arise from the unreliable speech recognition results under noisy conditions, the inherent ambiguity in the way humans express themselves in conversation, etc. The uncertainties in the four-level grounding state inference (channel, signal, intention and conversation) are modelled using Bayesian networks. The cost of grounding (grounding criterion) and subsequent cost of the grounding actions is modelled using decision networks (influence diagrams) that are essentially extended version of Bayesian networks. The authors have applied the method in three different dialogue systems - the Bayesian Receptionist (Horvitz and Paek, 1999), the Presenter (Paek et al., 2000), and the DeepListener (Horvitz and Paek, 2000)). The Receptionist is handling typical services offered by receptionists at Microsoft campus. For this purpose the system is able to detect a fixed number of user goals and map them to desired services. The presenter is a voice-driven presentation system that is able to detect only voice commands related to the slide manipulation. The DeepListener is a command and control system.

The model of Horvitz and Paek is influential in that it provides a computational model for unimodal grounding and error handling in a dialogue, based on identifying user goals and providing appropriate services. The authors give details on how such a system can be build by providing the Bayesian networks involved in the "Quartet" model. However, the intuition behind building the necessary topologies is not stated explicitly. The networks used seem monolithic, densely connected with multiple layers. Such type of Bayesian networks are difficult to interpret and reuse in other systems, since authors do not provide guidelines on how they were composed. Densely connected and multi-layered Bayesian networks are also known to be computationally expensive as far as probabilistic inference is concerned (Cooper, 1990; Jordan et al., 1999).

All of the presented error handling models based on grounding in human-computer dialogue are oriented towards extracting information mainly from one input modality, i.e. the speech modality. In human-robot interaction the speech modality can fail to provide sufficient information in order to avoid typical communication failures, such as the one resulting from a user that has abandoned conversation (Chapter 3). In noisy acoustic conditions the speech recognition can still process background noise and infer a valid user goal leading to "awkward" behavior from the side of the robot. In such conditions available modalities utilized by the robot for other purposes (e.g. navigation) such as laser and video provide additional information to be used in the grounding model. For example the lack of a user as detected in the laser scanner reading can point out at recognition errors that could otherwise result in valid user goals. The above observations outline the need for adapting and extending the initial states of the grounding model in Table 3.1 with new states associated with the different robot modalities (Chapter 7).

In the section that follows we investigate error handling techniques in voice-enabled interfaces of mobile service robots, focusing specially on error handling and the use of the concept of grounding.

3.4 Error handling in dialogue systems of service robots

The mobile tour-guide robot Jijo-2 (Matsui et al., 1999) used a microphone array for speech signal capture. The robot also required an explicit confirmation of each spoken input for more robust speech recognition. The potential of speech enhancement using microphone arrays has been exploited by other robots as well (Yamamoto et al., 2005; Choi et al., 2003; Hara et al., 2004).

The need for more systematic methods for error handling than the ad-hoc confirmation strategy of Jijo-2 is outlined in (Gieselmann and Waibel, 2005). In the study, the authors were using a simulated cooking robot that should answer questions related to cooking recipes input via keyboard. The goal was to investigate errors that can arise in dialogue with novice robot users, so that systematic clarification strategy can be designed. The authors turn attention to the fact that, in the case of system non-understanding, users try to provide shorter answers in attempt to contribute to the common ground on which mutual understanding between them and the robot can be build. In Section 3.4.1, we present real robot studies in which the problem of grounding in human-robot conversation is further discussed.

3.4.1 Grounding in human-robot interaction

In human-robot interaction, equipping the robot with the technical means for communication is not sufficient. It is also essential to answer the questions how should the communication proceed and how can the robot provide feedback about its state for the goal of efficient interaction with its user (Topp et al., 2004). The need of a systematic way of seeking and providing user feedback, during human-robot interaction is one of the main motivations behind the notion of grounding (Huttenrauch et al., 2004).

In (Huttenrauch et al., 2004) grounding, i.e. establishing common knowledge of a dialogue topic is seen as very important prerequisite for sustaining successful communication. In this study, grounding is defined at low and high levels of interacting. For high level grounding, the speech modality on the robot is used to extract information about the intention of the user. The robot in the study is a service robot, assisting a handicapped person in her/his everyday needs. In particular, the robot was designed to deliver objects to different locations (e.g. cups in the kitchen). The high level grounding is responsible for resolving ambiguities in user goal identification, when using natural spoken input to specify the robot tasks. The user goals can be related to one of two possible tasks (*Go to mission* and *Deliver mission*). Each of these tasks needs predefined pieces of information (e.g. location in the *Go to mission* location and object specification in the *Deliver mission*). Since some of the information could be missing or skipped in the spoken user input, grounding actions are used such as clarification questions to resolve the resulting ambiguity. The low level grounding on the other hand is dedicated to providing gestural feedback to the user through a small physical human-like character (CERO). The task specification (user goal) can be alternatively provided by means of a graphical user interface using a PDA.

The user goal specification in this study is based on natural spoken input. The input is interpreted using grammars and parsing into parameter/value pairs (e.g. location/kitchen, object/coffee) that define a given task. In other studies goals and subgoals are modelled using hierarchical Bayesian networks (Hong et al., 2005). Here goals and subgoals and related speech related features in a mixed initiative interaction with a service robot are related to variables in a Bayesian network. The network contains three levels of goal, subgoal and a feature level. If a goal is not inferred the subgoal level is monitored for inferring missing subgoals. The architecture can be used for grounding actions (clarification questions), although the process of grounding is not specified explicitly by the authors.

In (Aoyama and Shimomura, 2005) the robot is equipped with a layered attentional system that

is responding to high-level events related to interaction (e.g. missing concepts in conversation) as well as low-level events (e.g. high level of acoustic noise). The authors argue that combining low and high level feedback to the user about the state of the robot results in more intuitive human-robot interaction. A process similar to grounding is also discussed in (Sidner et al., 2004). The authors describe engagement rules in interaction with a static penguin-like robot Mel. They describe techniques very similar to the process of incremental grounding without explicitly referring existing work such as (Clark and Schaefer, 1989). Instead, they motivate their interactive engagement system from their user studies.

Both the studies (Hong et al., 2005) and (Huttenrauch et al., 2004) describe grounding from the perspective of high-level dialogue-system feedback-provision (goal clarification level). The grounding is performed using only the speech modality. However, the setting of human-robot interaction with mobile service robots differs from the more general case of human-computer interaction in that the user is free to move like the robot. User may also leave the robot at any time. Therefore, it is important that before providing high-level grounding actions the robot detects the state of the user attendance in the process of interaction.

3.4.2 Exploiting different input/output robot modalities

Detecting user activity is the purpose of the robot attentional system (Lang et al., 2003). This system can be seen as the component providing the robot with user and situation awareness. Situation awareness is the process that identifies entities in the surrounding environment that are essential for the process of human-robot interaction. For example, in a fetch and carry task the robot has to be aware of its user, of its current location and the location at which the robot needs to deliver. Attentional systems often employ multiple input modality information to achieve situation awareness.

In (Kleinehagenbrock et al., 2002) laser and video are used in the attentional system for detecting and tracking people in human-robot interaction using multimodal anchoring. The anchor is a high level description of the object of interest, i.e. the person. The multimodal anchor is composed of individual unimodal anchors containing attributes such as legs and face. The anchor is grounded with corresponding percepts from the laser scanner and video modality using a grounding relation. The grounding relation in this case is a set of rules that are used to find legs in the laser scanner reading and video skin-colored blobs in the video image. The method is used by a mobile service robot Biron, dedicated to tour-guiding.

In a follow-up study about the same robot (Li et al., 2005) the authors describe a multimodal (human-style) interaction system for the robot Biron, who has to learn new objects in the home of its user. The robot uses a multimodal interface based on speech and deictic pointing gestures for object specification. Grounding takes place on the higher interaction level of disambiguating the spoken input through clarification questions. In another recent study (Holzapfel and Gieselmann, 2004), the concept of grounding is used again in a multimodal (point and speak) human-robot interaction. Grounding is related to design clarification strategies, based on the so-called hold-on strategy. This strategy keeps the discourse information unchanged, although it might be inconsistent with the new user utterance. In this way a single incorrectly recognized utterance does not abort the current dialogue goal, but the user can still go on. In the study, the hold-on strategy is shown to improve human-robot interaction in problematic situations, arising for example in the case of misrecognized answers to clarification questions.

All the above studies concerning grounding in human-robot interaction are focused on high-level grounding in dialogue, relying on information derived from the speech modality. However, low-level grounding feedback from the side of the user, such as the state of attendance to the conversation,

can reveal very common situations that can produce recognition errors. Detecting the state of user attendance to the conversation would require additional information from modalities complementary to the speech modality. The attentional system of the robot can provide such information to the process of grounding in human-robot interaction. Finally, to enable low-level multimodal grounding, techniques for multimodal signal fusion need to be investigated. In the following section we present a brief review of existing techniques for multimodal signal fusion, outlining the use of Bayesian networks as a unifying statistical framework for modality fusion. In Chapter 6 and 7 of the thesis, we investigate further Bayesian networks for fusing multimodal information in recognition error handling techniques based on low-level grounding in conversations with mobile tour-guide robots.

3.4.3 Techniques for multimodal signal fusion

Multi-modal signal or sensor fusion is generally defined as any method that combines different signals to perform inferences that may not be possible from a single signal (Smith, 2003). Three basic fusion levels have been broadly defined in the literature: data-level fusion, feature-level fusion, and decision-level fusion. The methods used for inference depend on the particular task. Rule-based methods, fuzzy logic and neural networks have been used for tasks related to robotic navigation, as well as Kalman filtering (Kam et al., 1997).

Statistical Bayesian methods have been used for multi-target tracking as well (Smith and Srivastava, 2004; Bessière et al., 2003). The most common and simple Bayesian modelling for multimodal signal fusion can be represented by the following equation:

$$P(\Phi, S_1, S_2, ..., S_N) = P(\Phi)P(S_1|\Phi)P(S_2|\Phi)...P(S_N|\Phi). \tag{3.1}$$

In this equation Φ denotes the phenomenon of interest. The phenomenon under interest can be the location of the robot in the case of navigation, the coordinates of a target in 2D representation of a battlefield in the case of target tracking, or the user goal in the case of human robot interaction. $S_1, S_2, ..., S_N$ represent variables encoding the measured signal values from the different sensors. Equation 3.1 represent the joint probability function of the phenomenon and the different signal values measured by the sensors. This particular decomposition assumes that all the measured signal values are independent given Φ. In other words Φ is the cause and knowing the cause makes the consequences (the different sensor measurements) independent. The distributions $P(S_i|\Phi)$ are called the "sensor models". Knowing the joint distribution (Equation 3.1), it is a matter of simple normalization to obtain $P(\Phi|S_1, ...S_N)$ that is the query for the probability function over Φ given the sensor measurements:

$$P(\Phi|S_1, S_2, ..., S_N) \propto P(\Phi)P(S_1|\Phi)P(S_2|\Phi)...P(S_N|\Phi), \tag{3.2}$$

where $\propto$ is the symbol of proportionality. To choose a value for Φ different criteria such as MAP, posterior median, or MMSE criterium can be used. A detailed discussion on the appropriateness of each of the above optimality criteria can be found in (Smith, 2003).

Recently, Bayesian networks have emerged as a generalizing graph-based framework for creating statistical models (Jensen et al., 2002a; Russell and Norvig, 2003). A generic formalism known as Bayesian Programming (BP) that incorporates Bayesian networks as a structural unit has been offered as well. BP aims at representing a large class of general and more specific probabilistic models into a single modelling framework (Diard et al., 2003). Dynamic Bayesian networks have been already used to fuse sensor data in robotics for map building and localization (Shachter, 1998). Thorpe and McEliece (Thorpe and McEliece, 2002) have used Bayesian networks for scene analysis - combining many partial measurements from moving distributed sources (robots) in building a complete

scene. Bayesian networks have been also reported to improve the accuracy of phoneme recognition, when fusing the acoustic and video based phoneme counterparts in audio-visual recognition (Nefian et al., 2002).

Bayesian networks have been used for the purpose of classifier combination as well. Classifier combination problem can be reduced to the problem of multimodal signal fusion. In this way techniques for classifier combination can be used for the purpose of multimodal user goal identification in human-robot interaction. Conventional techniques for classifier combination work on score level, where the scores of the classifiers are combined using different rules, i.e. sum, product, median, weighted sum, etc. (Kittler et al., 1997; Kittler, 2000) to produce a more reliable score for the final classification. Bilmes and Kirchhoff (2000) have shown that all these rules can be generalized by Bayesian networks. Even more sophisticated combination rules that enhance the classification accuracy, can be achieved playing with the network topology, using predefined classes of Bayesian networks (Cheng and Greiner, 1999), as well as data-driven methods for deriving optimal topologies for the Bayesian network classifier (Pernkopf and Bilmes, 2005).

Experiments with Bayesian networks have been as well conducted in the domain of dialogue modelling for inferring dialogue acts (Keizer et al., 2002), anaphora referents and user goals in receptionist scenario (Horvitz and Paek, 1999). In our experiments we focus on inferring user goals in human-robot dialogues by fusing speech (acoustic-sensitive) with other acoustic-insensitive modalities. In order to define the user goals to be inferred and the inference method, a specification of the particular human-robot interaction is needed (Chapter 5).

3.5 Summary

In this chapter we have presented state of the art techniques for error handling in human-computer and human-robot dialogues. The focus has been on solutions that can be applied in the conditions of short-term human-robot interaction with a mobile tour-guide robot.

When applying recognition error prevention methods for robust speech recognition, a combination of techniques related to speech enhancement, robust feature extraction and methods related to manipulation of the recognition models can be used. We outline solutions that can be applied in the tour-guide interactive setting tailored to limited computational requirements.

In the second stage of detecting and correcting recognition errors, we outline the need for extending state-of-the art systematic error handling techniques to the interactive needs of tour-guide robots. Exploiting auxiliary information along with speech can lead to additional benefits when using speech recognition on inherently multimodal robotic platform.

Finally, to combine speech with additional non-speech modality information a multimodal signal fusion will be needed that can account for the uncertainties intrinsic to different modalities in the tour-guide interactive setting. Bayesian networks with their ability to generalize over a wide set of statistical models appear as an attractive tool for multimodal signal fusion.

Graphical models and decision theory

4

This chapter reviews theoretical elements about the two types of graphical models used in this thesis, i.e. Bayesian and decision networks. The goal of the chapter is to present the foundations behind Bayesian networks within the level of detail needed by the self-contained theoretical framework that can completely describe the models used later in this thesis. In our presentation, we take inspiration from Jensen (1996), presenting full proofs to the involved theorems, and correcting the mistakes that can be found in the original text.

Bayesian networks are a probabilistic framework for modelling problems that can be described by a set of causally related variables, where causality is not considered as deterministic. We define inference with Bayesian networks as the process of calculating *a posteriori* distribution over a set of unobserved variables of interest give another set of observed (evidential) variables. We then present a set of algorithms for inference with increasing sophistication and computational efficiency. The final algorithm presented is based on the "message passing" formalism for probability updating on a graphical structure known as junction tree that can be derived for each Bayesian network. The message passing algorithm also known as the junction tree algorithm can perform inference in time linearly dependant on the network nodes with Bayesian networks that have discrete random variables. We then show a particular type of network topologies in which continuous variables can be used with the same computational cost as in the discrete case inference using the message passing algorithm.

In order to perform consistent inference, Bayesian network conditional probability density (CPD) functions have to be learned from training examples. We present algorithms for CPD learning in the case when all or part of the network variables are observed.

We end the chapter showing a particular extension of the Bayesian networks, i.e. decision networks that can be used to address decision problems in which both the probability and utility of some state of the external environment is considered. This state is represented by a variable in the network. The decision on a final value for the state is governed by the principle of maximum expected utility, given the available evidence in the network and the preference of the system towards each state value.

4.1 Bayesian networks

4.1.1 Definition

Definition 2 (Bayesian network) *A Bayesian network is a directed graphical model defined by the triple (V, A, CPD), where V is a set of nodes associated to random variables, A is a set of directed arcs and CPD is a set of conditional probability distributions associated with the nodes' variables.*

Bayesian Networks (BNs) describe a joint probability distribution (pdf) over a finite set of random variables (Pavlovic, 1999). The joint pdf in the general case of N variables $(X_1, X_2, ...X_N)$ can be derived from the chain rule for the probabilities, after declaring the conditional independence assumptions given by the network's topology:

$$\begin{aligned} P(X_1, X_2, ..., X_N) &= \prod_{i=1}^{N} P(X_i|X_{i-1}, ..., X_1)) = \\ &= \prod_{i=1}^{N} P(X_i|Parents(X_i)) \end{aligned} \tag{4.1}$$

$Parents(X_i)$ are all the parent nodes for the node X_i, i.e. nodes that point to X_i. The equalities $P(X_i|X_{i-1}, ..., X_1) = P(X_i|Parents(X_i))$ declare the conditional independence assumptions encoded in the BN's graph. The BN encodes the particular pdf independence structure in the directed acyclic graph (DAG) in which nodes represent random variables, and the lack of arcs represents conditional independence assumptions between the variables. The directed arcs point from each parent variable to its dependent children variables. The variables and the dependencies among them can be used to model a specific process of interest that is random (stochastic) in its nature. The behavior of such a process can be fully described by the joint probability density function over all process variables. Later in the thesis, we use Bayesian networks to model processes related to voice-enabled interaction between human users and a tour-guide robot. The independence relations between the variables can greatly reduce the amount of information needed to parameterize a full joint pdf. In that sense a BN provides a compact representation through encoding the factorization of the joint pdf into independent CPD terms.

The BN's topology can be built on the basis of intuition (Jensen, 1996) drawn from the designers's knowledge about the process under modelling. There exist also methods for automatic BN structure learning. These methods require significant computational resources. Additionally, the approach depends heavily on sufficient amount of training data. The lack of such data may result in overfitting problems, i.e. models that fail to generalize well with unseen data. With these arguments in mind we focus in this thesis on constructing networks using knowledge and intuition.

Generally, the purpose of a BN is to give an estimate of the probabilities of unobserved events that would have allowed making decisions about the state of these events. Identifying such events, also called hypothesis events (Jensen, 1996), is the primary task when starting to build a BN model. The hypothesis events in the case of human-robot interaction can be related to particular user goals. For example, in the case of identifying user goals in the spoken interaction with a tour-guide robot, we can have two hypothesis events, i.e. "The user is willing to see the offered exhibit" and "The user is not interested". The hypothesis events are organized into a set of variables. A variable incorporates an exhaustive set of mutually exclusive events. In the presented example the two events can be organized into one variable UG (User Goal) with states 1 and 2 corresponding to the two possible user goals. Once the hypothesis events are organized in a set of variables, in order to estimate their

certainty some additional information that can provide evidence in favor of a particular hypothesis variable's state is needed. This is done by establishing information variables that can be discrete or continuous.

After defining the model variables the causal relations between them should be considered. These relations are represented by the arcs' direction and express a direct causal impact of one event on another. The arcs' directions do not necessarily coincide with the information flow direction. The arcs between the nodes point from all parent variables to their children variables. The intuition behind directionality represents the fact that the parent variables influence their children and this influence can be interpreted as a cause-effect relationship. This is why directed cycles are not allowed in BNs. From a probabilistic point of view the arcs converging at a given node specify the conjunction of all variables that appear as conditioning ones (parents) for the node's conditional probability distribution (CPD) term.

Probability tables represent the conditional probability distribution for a discrete variable. Arbitrary parametric CPDs can be assigned to the continuous ones. Conditional Gaussian distributions appear as a frequent choice in modelling continuous variables (Murphy, 2002). One reason for this option is that given some topological limitations (all arcs allowed except those pointing from a continuous parent to a discrete child) the resultant BN is a multivariate conditional Gaussian (Russell and Norvig, 2003). Gaussian mixture models (GMMs) appear as a special case of such a distribution in the case when the continuous variables have only discrete parents (Murphy, 2002). In such a case, if the independence assumptions in the network topology are supported by the variables' data used in training, the Bayesian network has the modelling power of the GMMs that can model arbitrary distributions. At the same time the Bayesian network encodes efficiently the parameter space of the model through the dependence assumptions and their cause/effect interpretation (Murphy, 2002; Russell and Norvig, 2003).

4.1.2 Properties

Three basic connections can exist in a general Bayesian network, e.g. serial, diverging and converging. These three variants are presented in Figure 4.1 (a).

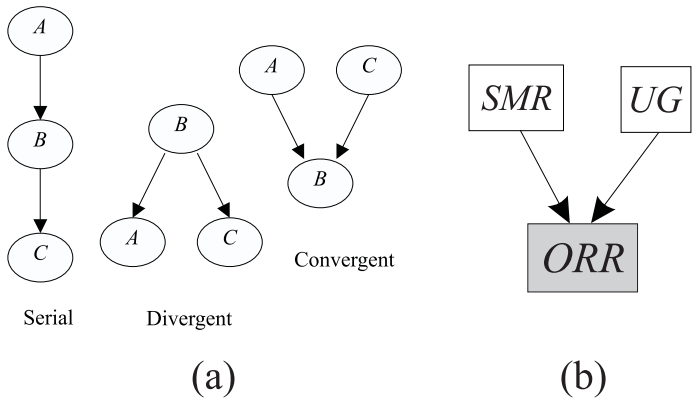


Figure 4.1: (a) Basic Bayesian network topologies, (b) example of convergent BN

The joint pdfs represented by these three Bayesian networks are as follows:

$$P(A,B,C) = P(A)P(B|C)P(C|B) = P(A)\frac{P(A|B)}{P(A)}P(C|B)P(B) = P(B)P(A|B)P(C|B) \quad (4.2)$$

$$P(A, B, C) = P(B)P(A|B)P(C|B) \tag{4.3}$$

$$P(A, B, C) = P(A)P(C)P(B|A, C) \tag{4.4}$$

From Equations 4.2 and 4.3 it is evident that the serial and divergent connections have equivalent independence properties among their variables.

We will now demonstrate the independence properties of the convergent connection, using a particular example. The Bayesian network depicted in Figure 4.1 (b) corresponds to the case of a converging connection. The set of network's variables $V = (UG, SMR, ORR)$ consists of three discrete variables (we depict discrete variable as squares, and observed variables are shaded in the example).

The event of particular *U*ser *G*oal in the voice-enabled interaction is associated with the variable UG. The user goals are generally associated to services that the service robot can provide, for example exhibit presentations in the case of the tour-guide robot. The event that the *O*bserved *R*ecognition *R*esult (ORR) can be unreliable is associated with the variable *S*peech *M*odality *R*eliability - SMR ($SMR = 1$ indicates reliable speech recognition, $SMR = 0$ indicates that speech recognition is unreliable). The variable ORR corresponds to observed recognition results that can be mapped to particular user goal values. The variables' conditional probability distributions (CPDs) are simply tables containing the values for the probabilities $P(UG)$, $P(SMR)$ and $P(ORR|UG, SMR)$. The joint pdf in this case can be written as:

$$P(V) = P(UG)P(SMR)P(ORR|UG, SMR).$$

The arcs in the graph can be seen as representing the causal relationships behind the variables in the above pdf. The two events (UG and SMR) can be seen as direct causes that can influence the particular value of ORR. Indeed if the current recognition result can be mapped into a user goal, our belief about the user goal being the cause for the particular ORR rises. If, we acquire additional evidence coming from the speech modality in favor of the $SMR = 0$ event (for example low signal-to-noise ratio), this new evidence will reduces our initial belief that UG is the cause for ORR, i.e. the observed speech recognition result, while increasing our belief that speech recognition is unreliable. The event $SMR = 0$ has explained away the observed recognition result and has lowered our initial belief in the observed speech recognition in noisy conditions. Such way of inter-causal relationship, commonly known as "the explaining away phenomenon" (Jensen, 1996) can be numerically encoded in the BN's CPDs and demonstrated using inference.

In order to determine if two variables are independent given some observed variables, we have to check if evidence can pass from one variable to another taking certain path in the Bayesian network. Evidence provided by an instantiated variable can pass through a serial or diverging connection, until the intermediate variable is not instantiated. In the case of convergent connection, the evidence in one of the parent nodes can affect the other only if their common child is instantiated. These properties are summarized by the rules of "*d*-separation", where *d* denotes directional. *d*-separation is a criterion from graph theory that accounts for the blocking of the flow of information between variables that are connected with arcs, independent from the direction of the arrows. *d*-separation can be used to infer local conditional independencies among the variables (Jensen, 1996). The *d*-separation rules state that (Jensen, 1996): Two variables A and C in a BN are *d*-separated if for all paths between A and C there is an intermediate variable B such that either the connection is serial or diverging and the state of B is known, or the connection is converging and neither B nor B's descendants have received evidence. If two variables are not *d*-separated, they are called

d-connected.

The "*d*-separation rules" are summarized in the so-called "Bayes ball" algorithm (Shachter, 1998) (Figure 4.2). The evidence in this case is regarded as a ball, entered at one variable and propagating in the network. The rules followed by the ball are illustrated in the figure. For example, in the case of Figure 4.2 (a) when the variable is not instantiated (blank) the ball is passing through the node (the two arrows indicating the direction of the evidence propagation are not blocked). When the variable is instantiated, the ball cannot pass (the two arrows are blocked).

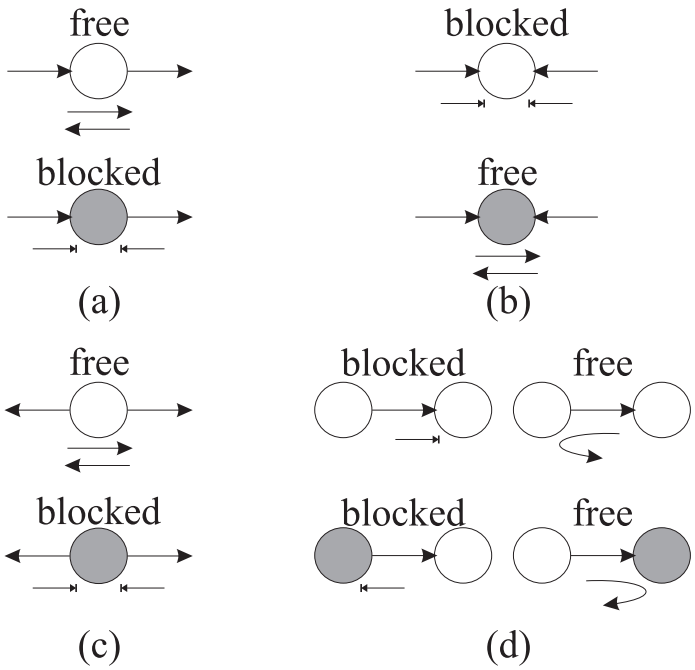


Figure 4.2: The "Bayes ball" algorithm: an evidence entered at some variable is seen as a ball bouncing in the network, between the variables whose conditional independence is of interest. If the ball can make its way from one variable to the other, the variables are dependent. The rules followed by the ball, while bouncing: (a) "Markov Chain" rule (serial connection), (b) "Competing explanations" rule (converging connection), (c) "Hidden variable" rule (divergent connection), (d) a boundary condition, when the ball hits the edge of the network

4.2 Inference in Bayesian networks

The basic task of probabilistic inference in Bayesian networks is to compute posterior distribution for a set of query variables, given some observed event, i.e. an evidence for some observed (evidential) variables. Formally, we calculate $P(X_Q|E)$, where $X_Q \in X$ is the subset of query variables from the full set of unobserved variables $X = \{X_0, ..., X_{L-1}\}; E = \{E_0, ..., E_{M-1}\}$ is the subset of the observed (evidential) variables and $V = X \cup E = \{V_0, ..., V_{N-1}\}$ is the set of all N random variables in the Bayesian network. Once the conditional probability distribution functions for all the nodes given their parents are defined, an exact or approximate inference on each node in the network can be done (Murphy, 2002; Pavlovic, 1999).

4.2.1 Exact inference by enumeration

In the simplest and least efficient case, exact inference can be performed through marginalizing the full joint pdf after entering the particular observed value (the evidence) for the observed variables $E = e$:

$$P(X_Q|E = e) = \alpha \cdot P(X_Q, E = e) = \alpha \cdot \sum_{X \setminus X_Q} P(V, E = e), \tag{4.5}$$

In the scope of this thesis we will be interested only in the case when X_Q is a discrete variable. In this case, $P(X_Q|E = e)$ denotes a posterior probability function over the possible values of X_Q. In the case of discrete variables, the posterior functions are probability tables. α is the normalization constant needed to make sure that the entries for $P(X_Q|E = e)$ sum up to 1. Note that taking into account the particular observed value ($E = e$) the term $\alpha = 1/P(E = e)$ remains constant for the set of values for X_Q and can be seen as a normalization constant. In that sense it is more efficient to use the already calculated $P(X_Q, E = e)$ values and normalize them, so that the sum of the final entries is 1 (Russell and Norvig, 2003). $X \setminus X_Q$ denotes set subtraction, i. e. the summation is over all possible values for the unobserved (non-evidential) variables that are in the set X and are not in the set X_Q. If all the BN variables are binary to compute $P(X_Q|E = e)$, we will need $O(2^N)$ operations (summations and multiplications) in total applying the simple enumeration method. Therefore, this method very soon becomes inefficient with large networks. In order to make inference more tractable, the calculation should make better use of already computed partial products and sums.

4.2.2 Inference by variable elimination

The enumeration algorithm can be improved by eliminating repeated calculations. The benefit comes at the cost of saving the result of partial calculations for later use. The simplest algorithm that makes use of partial calculation and can be used for inference in BNs is known as the variable elimination algorithm (Zhang and Poole, 1996). Variable elimination works by using the distributive law, evaluating expressions from right-to-left and storing intermediate results for later use.

Variable elimination operates on factorized joint density as in Equation 4.1. It takes as input a posterior distribution of interest, such as the query in Equation 4.5. The query defines a division of the full set of variables into a set of query variables, a set of evidential variables (variables that are fixed in inference) and a set of hidden variables, or the variables on which marginalization is performed. Given an elimination ordering π, the summations over the hidden variables are performed following π, taking only the hidden variables into account.

We will demonstrate the algorithm using the Bayesian network in Figure 4.3 (a).

The full joint pdf, encoded by the network is:

$$P(V) = P(D)P(C)P(A|D, C)P(B|A, C)P(F|D)P(G|A, F). \tag{4.6}$$

Let us assume that we are interested in calculating $P(X_Q|E) = P(C|B)$. Then, the posterior of interest can be written as follows:

$$\begin{aligned} P(C|E = \{b\}) &= \alpha \cdot \sum_{X \setminus X_Q} P(V, E = \{b\}) \\ &= \alpha \cdot \sum_{A,D,F,G} P(D)P(C)P(A|D, C)P(b|A, C)P(F|D)P(G|A, F), \end{aligned} \tag{4.7}$$

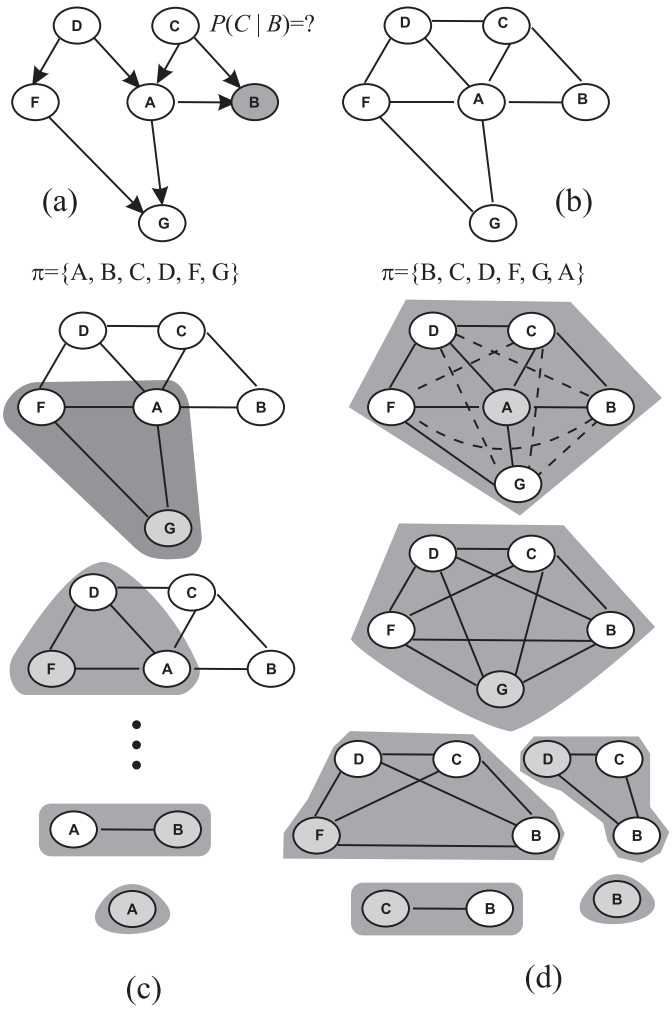


Figure 4.3: Bayesian network (a) its morilized graph (b), example of efficient node elimination (no fill-in arcs) (c), example for inefficient node elimination (d)

where $X = \{A, C, D, F, G\}$, $X \backslash X_Q = \{A, D, F, G\}$, $E = \{b\}$ is the particular assignment for the evidential variables E that stays fixed throughout the inference. Rearranging the conditional terms and using the distributive law, we move sum operators inside the product of conditional terms to get:

$$P(C|E = \{b\}) = \alpha \cdot \sum_{X \backslash X_Q} P(V, E = \{b\})$$

$$= \alpha \cdot P(C) \sum_{A} P(b|A, C) \sum_{D} P(D) P(A|D, C) \sum_{F} P(F|D) \sum_{G} P(G|A, F). \tag{4.8}$$

Every conditional term involved in Equation 4.8 can be seen as a factor over a set of variables. By a factor, we mean a non-negative function with an argument set over the variables involved in the initial conditional term. We write these factors as follows $\psi_1(C) = P(C), \psi_2(A, C) = P(b|A, C)$ ... $\psi_6(G, A, F) = P(G|A, F)$. Assuming that all variables in the BN are binary, the factor $\psi_6(G, A, F)$ is a $2 \times 2 \times 2$ probability table. The factor $\psi_2(A, C)$ is 2×2 table, since the first variable (B) in the conditional term is fixed to the particular evidence value. Thus, entering the particular evidence

value, has resulted in reducing the initial probability table ($P(B|A,C)$) into a table containing only the values for all possible configurations $\{B, A, C\}$, where B is fixed to $B = b$.

The calculation proceeds from right to left. The summation $\sum_G P(G|A,F)$ results in 1. The next summation $\sum_F P(F|D) \cdot 1$ sums up to 1 as well. Such factors can be initially removed from the query, since they do not contribute to the calculation. In general every variable that is not an ancestor of a query variable or an evidence variable does not affect the query (Russell and Norvig, 2003). We finally obtain the following expression for the posterior of interest:

$$\begin{aligned} P(C|E=\{b\}) &= \alpha \cdot \sum_{X \setminus X_Q} P(V, E=\{b\}) \\ &= \alpha \cdot P(C) \sum_A P(b|A,C) \sum_D P(D) P(A|D,C) \sum_F P(F|D) \sum_G P(G|A,F) \\ &= \alpha \cdot \psi_1(C) \sum_A \psi_2(A,C) \sum_D \psi_3(D) \psi_4(A,D,C). \end{aligned} \tag{4.9}$$

This expression is processed as follows. The right most sum contains two factors that have argument sets containing the variable D. They are multiplied using a point-wise product. The point-wise product leads to a new factor that we will call the pre-elimination factor for D, i.e. $\xi_D(A,D,C) = \psi_3(D) \cdot \psi_4(A,D,C)$. The argument set of the pre-elimination factor is formed out of the union of the argument sets of the factors involved in the product. Then, the product is performed using the corresponding elements that match the joint assignment, when restricted to the particular argument set of each of the multiplied factors. For example, if variables are binary, $\xi_D(A=t, D=f, C=t) = \psi_3(D=f) \cdot \psi_4(A=t, D=f, C=t)$. After computing the pre-elimination factor the corresponding variable (i.e. the variable D) is summed out (all entries indexed by the same A and C values but different D values are summed and D is dropped), resulting in the post-elimination factor $\xi^*(A,C) = \sum_D \xi_D(A,D,C)$. The post-elimination factor is used along with the other factors, in a similar way in the next summation, to get a pre-elimination and post-elimination factor. The recursive calculation is carried out until all variables are summed out. After performing the $P(C|B)$ computation for all values of C the result is normalized using the appropriate α.

The variable elimination algorithm can be written formally for the general case of N variables using a recursive definition. We first define the joint probability function as a product function over a set of factors $\psi_1, ..., \psi_N$ with argument sets $C_1, ..., C_N$:

$$P(V) = \alpha \cdot \prod_{n=1}^{N} \psi_n(C_n). \tag{4.10}$$

The factors $\psi_1, ..., \psi_N$ coincide with the particular conditional terms in the joint pdf factorization, and their argument sets $C_1, ..., C_N$ coincide with the union of the conditioned and conditioning variables in the particular conditional term.

Entering evidence

In the inference we are interested in calculating $P(X_Q|E=e)$, given an assignment of the evidential variables $E = e$. We can write $P(X_Q|E=e)$ as:

$$P(X_Q|E=e) = \frac{P(X_Q, E=e)}{P(E=e)}. \tag{4.11}$$

This equation can be written as:

$$P(X_Q|E=e) = \alpha \cdot P(X_Q, E=e) =$$
$$\alpha \cdot \sum_{V\setminus\{E\cup X_Q\}} P(V, E=e) = \alpha^{'} \cdot \sum_{V\setminus\{E\cup X_Q\}} \prod_{n=1}^{N} \psi_n^{'}(C_n^{'}), \tag{4.12}$$

where $\psi_n^{'}(C_n^{'})$ are the factors with instantiated evidence and argument sets $C_n^{'} = C_n \setminus E$. Hence, we can deal with evidence by simply instantiating it in all factors.

Variable elimination

After entering the evidence, we perform variable elimination for each variable X_k in the set $Y = X\setminus X_Q$ (the set of all hidden variables excluding the query variables). At every step of variable elimination, we rearrange the factorized pdf (Equation 4.12), dividing the product of factors in two groups. The first group contains all factors that do not include the variable to be eliminated (X_k), and the second group contains all factors involving X_k. Then, the product of factors (Equation 4.12) can be re-written as follows:

$$P(X_Q|E=e) = \alpha^{'} \cdot \sum_{Y\setminus X_k} \sum_{X_k} \prod_{n=1}^{N} \psi_n^{'}(C_n^{'}) \tag{4.13}$$
$$= \alpha \cdot \sum_{Y\setminus X_k} \Big(\prod_{s} \psi_s^{'}(C_s^{'}) \Big) \Big(\sum_{X_k} \prod_{l} \psi_l^{'}(C_l^{'}) \Big), \tag{4.14}$$

where s is the index over all factor domains that does not include X_k and l is the index over all factor domains that include X_k, i.e. $(1 \le s \le N, X_k \notin C_s)$ and $(1 \le l \le N, X_k \in C_l)$.

The right sum product is now computed explicitly resulting in the pre-elimination factor:

$$\xi_k(A_k) = \prod_{\substack{1\le l\le N \\ C_l^{'} \ni X_k}} \psi_l^{'}(C_l^{'}), \tag{4.15}$$

where

$$A_k = \bigcup_{\substack{1\le l\le N \\ C_l^{'} \ni X_k}} C_l^{'}. \tag{4.16}$$

The argument set of the pre-elimination factor is the union of the argument sets of all factors containing X_k. To get the post-elimination factor we sum-out over X_k:

$$\xi_k^*(A_k\setminus X_k) = \sum_{X_k} \xi_k(A_k). \tag{4.17}$$

It is clear that the number of computations performed by variable elimination depends on the chosen elimination order. Orderings that lead to pre-elimination factors with smaller argument sets will lead to a smaller number of computations. The significance of node elimination order can be intuitively demonstrated using a graphical interpretation of variable elimination.

Node elimination in Markov graphs

Each conditional term in Equation 4.7 is a function that can be associated with the corresponding variable and is called a factor. We will present the variable elimination algorithm from a graphical point of view. For that purpose we will first introduce the concept of Markov graphs.

Markov graphs are undirected graphs, also known as random Markov fields. They consists of nodes connected with undirected arcs. A Markov graph is composed out of special sub-graph units known as cliques. A clique in an undirected graph is a set of vertices C such that for every two vertices in C, there exists an edge connecting the two. This is equivalent to saying that the subgraph induced by C is a complete graph. The size of a clique is the number of vertices it contains. In a Markov graph every clique can be associated with a factor over a set of variables corresponding to the clique nodes. Then the formal definition for Markov graphs can be stated as follows

Definition 3 (Markov graphs) *Markov graphs are undirected graphs, defined as a triple* (V, C, Ψ)*, where* V *is a set nodes associated to random variables,* C *is a set of cliques and* Ψ *is a set of factors defined over the variables in the cliques.*

The factors associated with the graph cliques can be arbitrary non-negative functions. Such functions are also referred as potentials and they can include probability density functions as a special case. In this way a Markov graph can be also used to represent a factorized joint probability function, like a Bayesian network. Bayesian networks can be also converted into Markov graphs, using the process of graph moralization (Jensen, 1996). In this process the directionality of the graph is dropped and all parents for a given child node are connected with undirected arcs. Figure 4.3 (a,b) presents an example of a Bayesian network and its moralized Markov graph.

In the case of Markov graphs the independence statements are easy to interpret. A node is independent from the rest of the network, given its neighbors. Bayesian networks encode all conditional independence assumptions included in a Markov graph, however the inverse is not always true. Nevertheless, Markov graphs have proven to be very convenient graphical representations on which probabilistic inference can be done (Paskin, 2004).

After the process of moralization for each conditional term in the Bayesian network, there is an associated clique of nodes in the Markov graph. Thus the Markov graph clique structure can represent conveniently the argument sets of each conditional term included in the Bayesian network.

We now go back to the variable elimination algorithm. Let us assume that all variables from Equation 4.6 have to be eliminated using the order $\pi = \{A, B, C, D, F, G\}$. In the general case, the elimination is done only on a subset of the network variables, however the idea to be presented in this paragraph does not change if we have just a subset of the variables. Given the variable order π, the elimination is done backwards (first G than F, etc.), because the variable elimination is done from right to left. The elimination of a given variable X_k is done in two steps: compute pre-elimination factor and then sum-out to get the post-elimination factor. These two steps have their graphical correlates associated to particular manipulations performed on the Markov graph. Forming the pre-elimination factor on X_k creates an elimination clique (shaded region in Figure 4.3 (c,d)) over all neighbors of X_k, then summing out X_k to get the post-elimination factor, results in removal of the node and connecting the remaining neighboring nodes together to form the post-elimination clique. This sequence of operations on the Markov graph is known as the node elimination algorithm (Paskin, 2004). A node ordering π is chosen successfully if the node elimination results in small cliques and does not require fill-in edges. Figure 4.3 (c) presents one such example of node elimination using $\pi = \{A, B, C, D, F, G\}$. The size of the largest clique using this ordering is 3. However, using another node ordering (i.e. $\pi = \{B, C, D, F, G, A\}$) can result in much bigger cliques and many

fill-in edges, such as the example presented in Figure 4.3 (d). In this case, the largest clique size that dominates the number of computations is 6.

As demonstrated in the above example, node elimination allows to represent visually the complexity of variable elimination, depending on the elimination order. With special network topologies and good elimination order, variable elimination can perform inference with number of operations proportional to the number of network variables N, i.e. $O(N)$ number of operations. However, finding the best elimination order, in the general case is a $NP - complete$ problem. In practice, greedy solutions based on choosing the next variable in order to have a minimal clique size, work well (Kjaerulff, 1992).

4.2.3 The junction tree algorithm

The drawback of variable elimination is that it has to be run N times in the case of N different queries with the same evidence. Running variable elimination on each new query can result sometimes in redundant computations. This drawback is addressed in more powerful algorithms, based on the concept of message passing. These algorithm can perform inference in linear time, i.e. time or number of computations linearly dependant on the network size in terms of number of variables. The message passing algorithms operate on a graph structure known as the junction tree.

Message passing

The idea of message passing originates from the metaphor of distributed computing performed by several processors. In a distributed computing schema, processors are nodes that communicate over communication links, exchanging messages.

Definition 4 (Message passing) *Message passing can be seen as a process, in which each node in a network with N links to other nodes, waits until it gets a message on $N-1$ of its communication links to send a message on the remaining link.*

If the network of processors forms a tree, message passing is guaranteed to terminate after messages have been sent in both directions of each link. A tree is an undirected graph without cycles. The process of message passing in a tree is efficient, since it requires linear time in the number of nodes to terminate. Therefore, if the process of probabilistic inference can be seen as an instance of message passing, we have found an efficient way to perform inference. We now present how probabilistic inference can be reduced to the case of message passing.

Let us have a Markov graph representing a factorized form of a joint pdf $P(V)$:

$$P(V) = \alpha \cdot \prod_{i=1}^{N} P(X_i|Parents(X_i)) = \alpha \cdot \prod_{n=1}^{K} \psi_n(C_n). \tag{4.18}$$

Each clique C_i has a corresponding factor $\psi_i(C_i)$. Factors are non-negative functions and we associate them with the probability density functions (probability tables in the case of discrete variables). This is done in the following way. The conditional terms $P(X_i|Parrents(X_i))$ from the initial BN factored representation are associated with cliques C_i from the Markov graph, such that $(X_i \cup Parents(X_i)) \subseteq C_i$. One factor can be formed as a product of more than one conditional term, hence $N \geq K$. For example $P(A)$, $P(B|A)$ and $P(C|A,B)$ can be all associated with $C_i = \{ABC\}$.

Let $S_{ij} = C_i \cap C_j$ be the separator between cliques C_i and C_j. The nodes from the message passing network can be associated with the nodes in the Markov graph, i.e. the cliques C_i, while the links are associated to the separators S_{ij}. Message passing can be performed on a sub-class of Markov graphs that we will call cluster trees.

Definition 5 (Cluster tree) *A cluster tree over the set of variables V is a tree with nodes corresponding to clusters of variables from V. The union of all nodes is V.*

Cluster trees have two types of nodes, nodes corresponding to cliques (round nodes) and nodes corresponding to the intersection of the corresponding cliques or the separator (square nodes).

A BN and a cluster tree can represent the same pdf, if the following rules are followed when constructing the cluster tree:

- ♢ Form the clusters of variables, so that for each variable X_i there is at least one cluster node C_i that contains $X_i \cup Parents(X_i)$.
- ♢ Organize the nodes as a tree with separators $S_{ij} = C_i \cap C_j$ and initialize all separator factors $\phi(S_{ij})$ with value 1.
- ♢ Initialize all node factors $\psi(C_i)$ with the corresponding CPDs.

Now the joint pdf $P(V)$ represented by the cluster tree is formed as the product of all node factors divided by the factors over the separators. $P(V)$ coincides with the pdf represented by the BN as well, because all separator factors are initialized with 1s. Figure 4.4 depict an example cluster tree built for the BN in Figure 4.3.

The reason behind the use of separators is that when the node factor changes due to introducing new evidence, the product of all node divided by the separator factors stays invariant with respect to the operation used to propagate the new evidence in the network. The operation used in the junction tree algorithm to propagate evidence between nodes is called absorption.

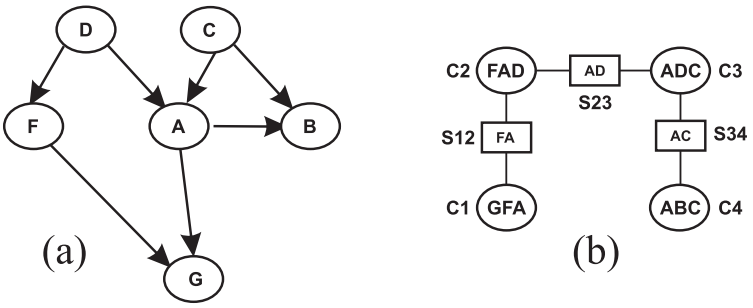


Figure 4.4: A Bayesian network (a) and a corresponding cluster tree (b)

We define absorbtion as follows (Jensen, 1996):

Definition 6 (Absorption) *Let C_i and C_j be neighbor cluster nodes in a cluster tree with intersection S_{ij}. The absorbtion of C_j from C_i is defined as:*

- ♢ *calculate* $\phi^*(S_{ij}) = \sum_{C_i \setminus S_{ij}} \psi(C_i)$
- ♢ *calculate* $\psi^*(C_j) = \psi(C_j) \frac{\phi^*(S_{ij})}{\phi(S_{ij})}$.

For discrete the variables, $\phi(\cdot)$ and $\psi(\cdot)$ correspond to tables and multiplication is point-wise as explained in Section 4.2.2. Division is performed in a similar manner, but we define $0/0 = 0$. After absorbtion we say that C_j has absorbed from C_i.

The main idea behind introducing absorption is to find an operation that is applied locally to each node and leads to global consistency of the joint pdf representation. The global consistency

means that the product of all node factors leads to the correct joint pdf. Since, both C_i and C_j contain the set S_{ij}, the following result should hold with a globally consistent representation:

$$\sum_{C_j \setminus S_{ij}} \psi(C_j) = \phi(S_{ij}) = \sum_{C_i \setminus S_{ij}} \psi(C_i). \tag{4.19}$$

If the relation in Equation 4.19 holds, we say that the link corresponding to S_{ij} is consistent or in other words the information that C_i and C_j hold about S_{ij} is the same. If all links are consistent, the tree is consistent. In this case, absorption does not have any effect on the factors of the cluster graph. Whenever new evidence about some node C_i arrives after C_j has absorbed from C_i all the three factors $\psi(C_j), \phi(S_{ij})$ and $\psi(C_i)$ hold the new information about the evidence. According to definition of absorption we can write then:

$$\frac{\psi^*(C_j)}{\phi^*(S_{ij})} = \frac{\psi(C_j)\frac{\phi^*(S_{ij})}{\phi(S_{ij})}}{\phi^*(S_{ij})} = \frac{\psi(C_j)}{\phi(S_{ij})}. \tag{4.20}$$

The last equation asserts that if we start with a BN, then we construct a cluster tree and perform a given number of absorptions, the cluster tree representation stays a representation of $P(V)$ that can be calculated as a product of all cluster factors divided by the product of all separator factors.

The question now remains how many absorptions to perform in order to be able to calculate $P(X_i)$ for a variable $X_i \in V$. Given the definition of absorption, we can prove the following theorem:

Theorem 1 (Consistency of message passing) *Let messages are passed in a supportive cluster tree T according to the message passing schema (Definition 4), then:*

(1) Messages can be passed until a message has been passed in both directions of each link.

(2) After full round of message passing T is consistent.

Proof: A supportive cluster tree is a tree in which $\psi(S_j)$ has zero entries that match the zero entries of $\psi(S_{ij})$. Without supportiveness absorption cannot be done properly. Supportiveness is guaranteed if we initialize all separator factors to tables of ones.

(1) Given a supportive cluster tree the case of a single node is trivial. We assert that the condition holds for the case of n nodes, and we want to show that then it also holds for the case of the tree T with $n+1$ nodes. Message passing is always started with a leaf node, e.g. X_a sending message to its single neighbor X_b. Let us assume that message passing has been performed until a moment that not all links have received a message in both directions. Consider the tree $T \setminus X_a$ that has n nodes. Let us assume that a message from X_a has been passed. According to the induction hypothesis, message cannot be passed in $T \setminus X_a$ any more after the messages have been sent in both directions of each link in $T \setminus X_a$. In that case a legal message can be sent from X_b to X_a.

(2) If we have a single node tree, the theorem is true. With more nodes, let us assume that an arbitrary link (C_i, C_j) with a separator S_{ij} has been traversed by a message from C_j to C_i. The original factors associated with the link are $\psi(C_j)$, $\phi(S_{ij})$ and $\psi(C_i)$. After passing a message from C_j to C_i we get $\phi^*(S_{ij}) = \sum_{C_j \setminus S_{ij}} \psi(C_j)$. Next time when a message has to pass from C_i to C_j the factors $\phi(S_{ij})$ and $\psi(C_j)$ are still not changed. C_j has not received any other messages, because we have a tree, and also because of (1.). The factor of C_i now is $\psi^{**}(C_i)$. After passing a message we have:

$$\phi^{**}(S_{ij}) = \sum_{C_i \setminus S_{ij}} \psi^{**}(C_i), \text{ where } \psi^{**}(C_j) = \psi(C_j)\frac{\phi^{**}(S_{ij})}{\phi^*(S_{ij})}.$$

We can write then:

$$\sum_{C_j \setminus S_{ij}} \psi^{**}(C_j) = \sum_{C_j \setminus S_{ij}} \psi(C_j) \frac{\phi^{**}(S_{ij})}{\phi^{*}(S_{ij})} = \frac{\phi^{**}(S_{ij})}{\phi^{*}(S_{ij})} \sum_{C_j \setminus S_{ij}} \psi(C_j)$$
$$= \frac{\phi^{**}(S_{ij})}{\phi^{*}(S_{ij})} \phi^{*}(S_{ij}) = \phi^{**}(S_{ij}) = \sum_{C_i \setminus S_{ij}} \psi^{**}(C_i).$$

Hence the links and the tree are consistent. □

Message passing on junction trees

Although message passing results in a consistent cluster tree, this does not guarantee global consistence. We say that a cluster tree is globally consistent, if for any two nodes C_i and C_j with a intersection I we have

$$\sum_{C_j \setminus I} \psi^{**}(C_j) = \sum_{C_i \setminus I} \psi(C_i).$$

The above equality holds when C_i and C_j are neighbors, however it does not hold in the general case. Consistence implies global consistence only in a special class of cluster trees, i.e. junction trees. Consistence of cluster trees does not imply global consistence, since a variable X_i can be put in two locations in the tree such that information on X_i cannot be passed between the two locations. To overcome this problem we restrict the cluster trees to the class of trees in which all nodes on the path between each two nodes contain their intersection.

Definition 7 (Junction tree) *A junction tree is a cluster tree in which for any pair of nodes C_i and C_j, all nodes on the path between C_i and C_j contain their intersection $C_i \cap C_j$.*

Theorem 2 (Consistency of junction trees) *A consistent junction tree is globally consistent.*

Proof: Let C_i and C_j are nodes in a locally consistent junction tree, and I is their intersection. According to Definition 7, I is a subset of all nodes on the path between C_i and C_j. Since the tree is locally consistent the marginal probability on I is the same for all nodes in the path. □

We will demonstrate that if we construct a junction tree that corresponds to a Byesian network, we can have an efficient algorithm for inserting evidence and probability updating. We first show that:

$$\psi(C_i) = \sum_{V \setminus C_i} \Psi(V), \tag{4.21}$$

where $\Psi(V)$ is the product of all node factors divided by the separator factors of a consistent junction tree over V.

Proof: To prove the above equality we use again mathematical induction. It clearly holds in the case of a single variable tree T. We assert that it holds for a tree $T^{'}$ with n variables and we prove that it holds for the tree T with $n+1$ variables. Let C_i be a leaf in T connected to the node C_j and S_{ij} be the separator. We remove C_i to get $T^{'}$ with set of nodes $V^{'}$. Then by definition $\Psi(V) = \Psi(V^{'}) \frac{\psi(C_i)}{\phi(S_{ij})}$, where $\Psi(V^{'})$ is the product of all node and separator factors in $T^{'}$. Let $D = C_i \setminus S_{ij}$ and $H = C_j \setminus S_{ij}$. From the junction tree property it follows that $D \cap V^{'} = \emptyset$. Since T is consistent, it follows that:

$$\sum_{D} \psi(C_i) = \phi(S_{ij}) = \sum_{H} \psi(C_j).$$

Then:

$$\sum_{D} \Psi(V) = \sum_{D} \Psi(V^{'}) \frac{\psi(C_i)}{\phi(S_{ij})} = \frac{\Psi(V^{'})}{\phi(S_{ij})} \sum_{D} \psi(C_i) = \Psi(V^{'}) \frac{\phi(S_{ij})}{\phi(S_{ij})} = \Psi(V^{'}). \tag{4.22}$$

By Equation 4.22 and the induction hypothesis we have:

$$\sum_{V \setminus C_k} \Psi(V) = \sum_{V' \setminus C_k} \sum_{C_i \setminus S_{ij}} \Psi(V) = \sum_{V' \setminus C_k} \sum_{D} \Psi(V) = \sum_{V' \setminus C_k} \Psi(V^{'}) = \psi(C_k) \tag{4.23}$$

for all C_k in $T^{'}$. It remains only the case of the node C_i. For this case we have:

$$\sum_{V \setminus C_i} \Psi(V) = \sum_{V' \setminus S_{ij}} \Psi(V^{'}) \frac{\psi(C_i)}{\phi(S_{ij})} = \frac{\psi(C_i)}{\phi(S_{ij})} \sum_{V' \setminus S_{ij}} \Psi(V^{'})$$

$$= \frac{\psi(C_i)}{\phi(S_{ij})} \sum_{C_j \setminus S_{ij}} \sum_{V' \setminus C_j} \Psi(V^{'}) = \frac{\psi(C_i)}{\phi(S_{ij})} \sum_{C_j \setminus S_{ij}} \psi(C_j) = \frac{\psi(C_i)}{\phi(S_{ij})} \phi(S_{ij}) = \psi(C_i). \tag{4.24}$$

□

We are now ready to prove the following theorem:

Theorem 3 (Correctness of message passing) *Let BN be a Bayesian network representing $P(V)$, and let T be a junction tree corresponding to BN. Let $e = \{e_1, e_2, ..., e_m\}$ be findings on the evidential variables $\{E_1, E_2, ..., E_m\}$. For each i find a node E_i and enter the evidence into its corresponding factor. Then, after a full round of message passing we have for each node C_i and separator S_{ij}:*

$$\psi(C_i) = P(C_i, e),\ \phi(S_{ij}) = P(S_{ij}, e) \text{ and } P(e) = \sum_{C_i} \psi(C_i).$$

Proof: After entering the evidence into each node factor (Section 4.2.2), $P(V, e)$ can be formed as product of the initial node factors divided by the product of the initialized separator factors. According to Theorem 1, after a full round of message passing T is consistent, and $P(V, e)$ is the product of all node factors divided by the separator factors. Then according to Theorem 2 and Equality 4.21, we can write that $\psi(C_i) = \sum_{V \setminus C_i} P(V, e) = P(C_i, e)$ and $\phi(S_{ij}) = \sum_{C_i \setminus S_{ij}} \psi(C_i) = P(S_{ij}, e)$. Finally $P(e) = \sum_{C_i} P(C_i, e) = \sum_{C_i} \psi(C_i)$.

□

Constructing junction trees

To complete the description of the junction tree algorithm we have to describe how a junction tree can be constructed from the original BN. A junction tree can be constructed from a triangulated moral graph corresponding to the initial Bayesian network. In a triangulated undirected graph any cycle of length > 3 has a chord. In Section 4.2.2, we have described how to derive a moral graph from a BN. A triangulated graph can be derived from a moralized graph after applying the node elimination algorithm (Section 4.2.2) on the moral graph with a predefined elimination order. Finally, a graph is triangulated if and only if all of its nodes can be eliminated one by one without adding any fill-in arcs. Note that there are several triangulations of the graph, depending on the elimination order. Intuitively, triangulations with as few fill-ins as possible are preferred.

Definition 8 (Junction graph) *A junction graph for a undirected graph G is an undirected, labelled graph. The nodes are the cliques in G. Every pair of nodes with a non-empty intersection has link labelled by the intersection*

To identify cliques in a triangulated graph G, we can use the following heuristic. Let $\{X_1, ..., X_N\}$ be an elimination sequence for G, and let C_i be the set of variables containing X_i and all its neighbors at the time of elimination. Then every clique of G is a C_i for some i.

A junction tree can be derived as a spanning tree of a junction graph. A spanning tree is a subtree of a graph that includes all the graph nodes. A spanning tree is a junction tree if it has the property that for each pair of nodes, C_i and C_j, all nodes on the path between C_i and C_j contain their intersection $C_i \cap C_j$. In the literature of Bayesian networks this property is also known as the Running Intersection Property (RIP).

Theorem 4 (Existence of a junction tree) *An undirected graph is triangulated if and only if its junction graph has a junction tree.*

Proof: We have to prove the following statements:
(1) A connected undirected graph is triangulated if it has a junction tree.
(2) Any connected triangulated graph has a junction tree.
(1) Induction in the number of nodes. The condition (1) is true for the case of two nodes. We assert that it is true for all graphs with less than n nodes. Then, let G be a connected graph with n nodes, and let T be a junction tree for G. Since T is a tree, there is a clique C with only one neighbor $C^{'}$ in T. Let A is a node belonging to $C \backslash C^{'}$. A can only be a member of the clique C, because of the fact that T is a junction tree. Then all neighbors of A are in C and are therefore pairwise linked.

If we remove A from C the graph is reduced to $n-1$ nodes. If the new clique after removing A becomes a subset of $C^{'}$ we remove C from T. The junction tree T^* after removing A is a junction tree for G^*. According to the induction hypothesis, G^* is triangulated, and therefore G is also triangulated.

(2) Induction in the number of nodes. The condition (2) is obviously true for the case of two nodes. We assert that it is true for all graphs with less than n number of nodes. Let G is a triangulated graph with n nodes. Since a triangulated graph can be seen as a result after applying node elimination, there is at least one node A pairwise connected to all its neighbors in a clique C. G and G^* resulting from removing node A have the same cliques except C. The corresponding clique in G^* could be $C \backslash A$. Since G is triangulated, G^* is also triangulated. By the induction hypothesis G^* has a junction tree T^*. Now we construct the tree T out of T^* according to the following rules:

- ♢ If $S = C \backslash A$ is a clique in T^* add A to S.
- ♢ If $S = C \backslash A$ in not a clique in T^* then $S \in C^{'}$, where $C^{'}$ is a clique in T^*. Then add C as clique with a link S to $C^{'}$.

Using Definition 8 it is easy to verify that the resulting tree T is a junction tree for G. □

The above theorem shows that a junction can always be derived from a triangulated graph. Since some of the cliques in the triangulated graph can be subsets of others, it is conventional to form a junction graph from maximal cliques removing the redundant ones.

Finally, to construct a junction tree we can use the individual link weights in junction graph with maximal cliques. The link weight is equal to the number of variables in the label of the link. In other words, the link weight for two nodes C_i and C_j is the number of variables they have in common, i.e. $|C_i \cap C_j|$. The weight of the whole junction tree is the sum of the individual link weights. Then, any maximal weigh spanning tree of the junction graph is a junction tree (see (Aji and McEliece, 2000) for a proof). The above statement provides an easy way for construction of junction trees: choose successively a link of maximal weight unless it creates a cycle, which is known as the Kruskal's algorithm.

In conclusion, junction trees are constructed from a BN through the following consecutive steps: (1) Moralization, (2) Triangulation and forming of maximal cliques junction graph, (3) Applying the Kruskal's algorithm to construct a junction tree.

The most problematic step in the junction tree formation is the triangulation step. Any elimination can produce triangulation, however as with variable elimination the size of the resulting cliques may be intractable in terms of space and subsequent computation requirements, needed by inference using message passing. Like in the case of variable elimination, greedy heuristics can be applied in the triangulation step to ensure sufficiently small final cliques in practice.

4.2.4 Message passing with continuous variables

In the previous sections we have defined the message passing algorithm only for discrete variables. In the case of hybrid Bayesian networks (including both discrete and continuous variables) the factors corresponding to continuous variables typically represent parametric models of distributions. The exponential family of distributions, in particular Gaussian distribution are often used in Bayesian networks, resulting in the so-called Linear Continuous Gaussian (LCG) networks. In these networks arcs are allowed to point from discrete to continuous variables, but not vise-versa. Then, we end up with a conditional linear Gaussian distributions for the continuous nodes. For every configuration of the discrete parents the continuous node is weighted linear sum of its continuous parents and some Gaussian noise.

We define a sub-class of the LCG Bayesian networks that will be used throughout this thesis. In this class, additionally we allow only discrete parents and we set that all continuous variables have to be observed. We are interested only in discrete unobserved variables in the thesis, while continuous variables will be associated to observed features. Many practical problems, related to e.g. pattern recognition, classification, etc. make use of such a problem definition. In this case the resulting pdf is a mixture of Gaussians. Each continuous node Y_i is represented in the Bayesian network with a table of possibly multivariate means and variances $(\mu_j, \sigma_j^2)_{1 \leq j \leq M}$, one for each possible configuration $j = \{1, 2, ..., M\}$ for the discrete parents $Parents(Y_i)$.

In order to reduce the case of CLG to the case of mixtures of Gaussians we can unite the continuous nodes with their parent continuous nodes resulting in a single multivariate Gaussian node.

To allow the use of the discrete message passing algorithm with continuous variables, when entering the evidence, we can calculate the probability for the particular observed value $Y_i = y$ in all possible configurations for the parent variables. In this way' we form the table $(P(Y_i = y|\mu_j, \sigma_j^2))_{1 \leq j \leq M}$ and use this table in the factor that includes the particular $P(Y_i|Parents(Y_i))$ term. In that way, we can use the message passing algorithm described in Section 4.2.3, without the need to define multiplication and division for continuous Gaussian factors. In this case, we can take advantage of the linear complexity of inference with message passing in the number of network nodes.

4.2.5 Complexity of inference

The time complexity of exact inference in Bayesian networks with linear conditional Gaussians is NP hard (Murphy, 2002; Lerner and Parr, 2001). The "junction tree" algorithm used for inference in our case is done in two phases, e.g. constructing a junction tree from the original Bayesian network and performing inference on a junction three after entering the evidence. The NP-hardness comes into place when the junction tree CPDs are constructed (Russell and Norvig, 2003). In our case, we have a static Bayesian network, i.e. its topology remains unchanged during the different inference

instances and the junction tree CPDs need to be constructed only once. In addition, our continuous variables are observed, which avoids the problem of marginalizing continuous variables. Thus the time of exact inference, once the junction tree is constructed, is linearly dependent on the number of network nodes.

4.3 Bayesian network CPD Learning

In order to perform consistent inference, estimates for the conditional probability distribution parameters have to be learned from training examples for the network variables.

The goal of the CPD parameter learning is to obtain estimates for the conditional distribution functions of the variables from data (the conditional probability tables for the discrete variables and the parameters of the Gaussian pdfs for the continuous ones).

4.3.1 Full observability

In the case of full observability of the variables in the training set, the estimation can be done with random initialization and a Maximum Likelihood (ML) training technique. During the training, the CPD parameters are adjusted in order to maximize the likelihood of the model with respect to the training data examples (Appendix C.2 in (Murphy, 2002)). The likelihood computation formulae needed to train the Bayesian networks used in our experiments are given below.

The likelihood of a Bayesian network, defined over a graph G is given by the formula:

$$L = \log \prod_{m=1}^{M} P(D_m|G) = \sum_{i=1}^{N} \sum_{m=1}^{M} \log P(X_i|Parents(X_i), D_m), \tag{4.25}$$

where $D = \{D_1, D_2, ..., D_m, ..., D_M\}$ is the set of training examples (cases), containing the values for the N variables in the network G. $Parents(X_i)$ are the parents for each node variable X_i. This likelihood function is decomposed into terms one for each node corresponding to the node's CPD. We need to specify the log-likelihood for discrete and Gaussian CPDs given the subset of their local training data. In the case of discrete - tabular CPD the likelihood is $L = \sum_{i,m} \log \prod_{j,k} P(X_i = k|Pa(X_i) = j))^{I_{ijkm}}$, where I_{ijkm} is the indicator function of the event $(X_i = k, Parents(X_i) = j)$ in the case of D_m. It can be shown using derivatives and Lagrange multipliers that:

$$P(X_i = k|Pa(X_i) = j) = \frac{N_{ijk}}{\sum_{k'} N_{ijk'}}, \tag{4.26}$$

where N_{ijk} is the number of times the event $(X_i = k, Pa(X_i) = j)$ occurs in the training data. And the likelihood becomes:

$$L = \sum_{ijk} N_{ijk} \log \frac{N_{ijk}}{\sum_{k'} N_{ijk'}} \tag{4.27}$$

For the case of the continuous variables, we have Gaussian nodes with discrete parents. In such a case, when the parents are hidden the continuous CPD is in fact a mixture of Gaussians. The log-likelihood for a continuous node Y is then given by the formula: $\log \prod_{m=1}^{M} \prod_{i=1}^{K} [N(y_m|\mu_i, \Sigma_i, D_m)]^{q_m^i}$, where y_m is the vector of continuous values in the case D_m, K is the number of possible discrete parents' configurations, and q_m^i is the indicator of the event $(Pa(Y) = i|D_m)$. The formulas for

calculating the means and variances for the K possible parents' configurations are given below:

$$\mu_i = \frac{\sum_m q_m^i \cdot y_m}{\sum_m q_m^i} \tag{4.28}$$

$$\Sigma_i = \frac{\sum_m q_m^i \cdot y_m y_m^T}{\sum_m q_m^i} - \mu_i \mu_i^T \tag{4.29}$$

Proofs for the above formulas can be derived based on Appendix C in (Murphy, 2002).

4.3.2 Partial observability

In the case of partial observability of variables during learning of the CPDs the log-likelihood is:

$$L = \sum_m \log(P(D_m)) = \sum_m \log \sum_h P(H = h, V \backslash H = D_m), \tag{4.30}$$

where H is the set of the hidden variables, and $V \backslash H$ is the set of observed variables which take on a value D_m. Unlike the fully observed case, the log-likelihood L cannot be decomposed into a sum of local terms one per node. Generally, there are two approaches to perform CPD learning with hidden variables, i.e. gradient ascent and Expectation Maximization (EM). As shown in (Murphy, 2002) the gradient ascent is very similar to EM. Therefore we concentrate on the EM algorithm.

The EM algorithm

The EM basic idea is to use Jensen's inequality (Cover and Thomas, 1991) to get a lower bound on the log-likelihood and to maximize this bound through a series of iterations. Jensen's inequality states that for any concave function f, we have:

$$f\left(\sum_j \lambda_j y_j\right) \geq \sum_j \lambda_j f(y_j), \tag{4.31}$$

where $\sum_j \lambda_j = 1$. Since the log-likelihood function is concave, we can use Jensen's inequality to get:

$$\begin{aligned} L = \sum_m \log \sum_h P(H = h, D_m) &= \sum_m \log \sum_h q(h|D_m) \frac{P_\theta(H = h, D_m)}{q(h|D_m)} \\ &\geq \sum_m \sum_h q(h|D_m) \log \frac{P_\theta(H = h, D_m)}{q(h|D_m)} \\ = \sum_m \sum_h q(h|D_m) \log(P_\theta(H = h, D_m)) &- \sum_m \sum_h q(h|D_m) \log(q(h|D_m)), \end{aligned}$$

where the function q has to satisfy the following conditions: $\sum_h q(h|D_m) = 1$ and $0 \leq q(h|D_m) \leq 1$. Maximizing the lower bound with respect to q results in $q(h|D_m) = P_\theta(h|D_m)$. This is called the Expectation step (E step), and makes the bound tight.

Maximizing the lower bound with respect to the free parameter θ' is equivalent to maximizing the expected complete-data log-likelihood:

$$E_q[l_c(\theta')] = \sum_m \sum_h q(h|D_m) \log(P_{\theta'}(H = h, D_m))$$

This is called the maximization step (M step). This step is efficient if the corresponding complete-data problem is tractable, and q has a tractable form.

If $q(h|D_m) = (P_\theta(h|D_m))$, as in the exact EM case, then the Equation 4.32 is often written as:

$$Q(\theta^{'}|\theta) = \sum_m \sum_h P(h|D_m, \theta) \log(P(h, D_m|\theta'))$$

In (Dempster et al., 1977) it is proven that choosing $\theta^{'}$ such that $Q(\theta^{'}|\theta) > Q(\theta|\theta)$ is guaranteed to ensure $P(D|\theta^{'}) > P(D|\theta)$, i.e. increasing the expected complete data log-likelihood will increase the actual (partial) data log-likelihood. This is because using $q(h|D_m) = (P_\theta(h|D_m))$ in the E step makes the lower bound touch the actual log-likelihood curve, so raising the lower bound at this point will also raise the actual log-likelihood curve.

In the case of multinomial CPDs, the expected complete-data log-likelihood becomes:

$$Q(\theta^{'}|\theta) = \sum_{ijk} E[N_{ijk}] \log(\theta^{'}_{ijk}), \tag{4.32}$$

where $E[N_{ijk}] = \sum_m P(X_i = k, Pa(X_i) = j|D_m, \theta)$, so the M step, where $\theta = argmax_{\theta'}(Q(\theta^{'}|\theta))$, becomes:

$$\hat{\theta}_{ijk} = \frac{E[N_{ijk}]}{\sum\limits_{k'} E[N_{ijk'}]}. \tag{4.33}$$

This is a generalization of the EM algorithm for HMMs. The idea of the algorithm can be applied to any BN (Lauritzen, 1995). The two basic steps are: compute the expected sufficient statistics, using an inference algorithm (compute $\sum_m P(X_i = k, Pa(X_i) = j|D_m, \theta_{old})$); use these statistics in the M step as if they were actually sufficient statistics computed from data and repeat.

4.4 Decision theory

The inferred posterior distribution $P(X_Q|Y)$ for the query variable X_Q, can be used for making decisions on a particular value for X_Q, based on the observed evidence $E = e$. If X_Q is a discrete variable this last step can be seen as a classification problem in which X_Q is the classification variable. Different optimality criteria for assigning X_Q to one of its possible class values exist. To select the most likely X_Q we use an $argmax$ criterion:

$$\hat{x_q} = \arg\max_{x_q}(P(X_Q = x_q|E = e)) \tag{4.34}$$

In order to include preference towards a given state of the X_Q variable, decisions can be based on principles from utility theory.

4.4.1 Utility theory

The principle of maximum expected utility (MEU) is used by modern decision theory and artificial intelligence for modelling the process of decision-making or the strategy of action selection of a utility-driven agent (Russell and Norvig, 2003).

Figure 4.5 depicts the architecture for a utility-driven agent. Such an agent maintains an internal state representation of its environment given its sensors' information. A utility function is used to model the agent's preferences for the different actions through which the agent can manipulate its environment given its internal state. The utility function assigns a numerical value to each

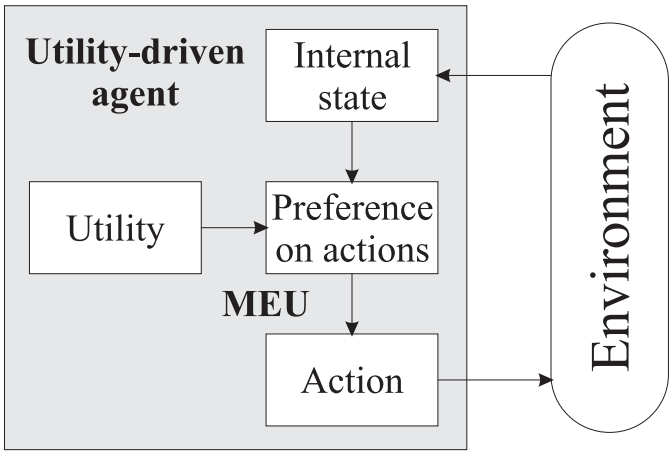


Figure 4.5: Architecture for a utility-driven agent

agent's actions, given the current state of the environment. Finally, the process of action selection is modelled by combining principles from probability and utility theories. Probability theory is used to model the agent's internal state, given the information (evidence) extracted from its sensors. Utility theory is used to model the agent's preferences between the states of the external environment resulting from a taken decision (executed action). These preferences are captured by the utility function as mentioned above. We use utility function $U(s, a)$ to denote the utility of an action, given that the agent is in a state s. $P(S = s|E = e)$ will denote the probability of each state value, given the current evidence $E = e$ from the sensor data. Then the maximum expected utility is defined by the following equation (Jensen, 1996):

$$MEU(\hat{a}|e) = \arg\max_{a} \sum_{s} P(S = s|E = e) \cdot U(s, a) \tag{4.35}$$

The maximum expected utility principle in decision theory states that an intelligent agent should choose the action that maximizes the expected utility of that action, given the sensor evidence for and the state of the world at the instance of decision-making. This kind of utility driven decisions can be visually represented and implemented with the help of decision networks (Russell and Norvig, 2003; Paek and Horvitz, 2003).

4.4.2 Decision networks

In a decision network (DN) there are three types of nodes: chance nodes (ovals), decision nodes (rectangles) and utility nodes (diamonds). An example of a decision network is shown in Figure 4.6. The chance nodes represent random variables. These variables are similar to the BN variables. The agent is usually uncertain about the exact values of these variables. Some of the chance nodes can represent features extracted from the agent sensors; others can represent different aspects of the agent's internal state. In the example presented in Figure 4.6 the chance node represents the state of the user goal variable in human-robot interaction. Decision nodes represent possible choice of actions. In the example presented in Figure 4.6 the decision node incorporates the possible dialogue continuations based on particular UG states. The utility nodes represent the utility function. Since the utility function depends on the agent's internal state and the actions, utility nodes usually have one or more chance nodes and the decision node as parents. Bayesian networks (Russell and Norvig,

2003; Jensen, 1996) are often used to model the probabilistic dependencies between the chance nodes and serve as an input to the decision network.

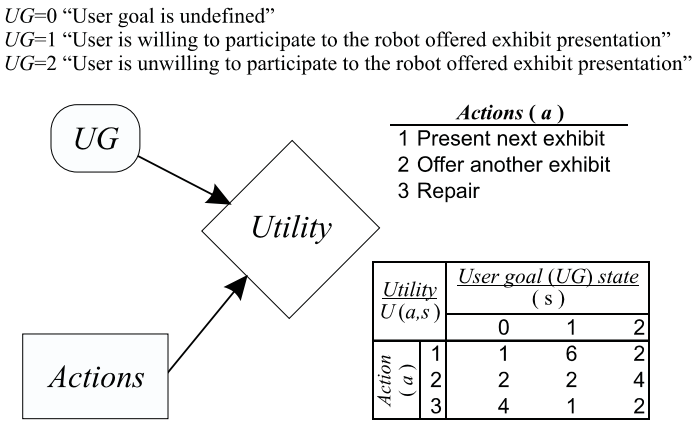


Figure 4.6: Example of decision network

Bayesian networks specify a family of statistical models, equipped with a unified set of efficient algorithms for inference (Jensen, 1996), e. g. computing posterior probability over set of "query variables", given an assignment for some set of observed variables in the network. The observed variables are usually named evidence variables. Therefore, Bayesian networks can be used to produce the probability values on the state variables, i. e. $P(S = s|E)$ for the utility-driven agent. Then applying Equation 4.35 will result in selecting the action with MEU, given the set of possible actions.

4.5 Summary

In this chapter, Bayesian and decision networks were presented. The main goal of a probabilistic model represented by a Bayesian network is to decide on a value of hidden variable of interest given observed evidential variables. This decision require an inference process which in the general case is NP-hard.

We presented algorithms for inference that increase in their theoretical sophistication, while taking advantage of the local BN structure to perform efficient inference. In particular, we presented an algorithm (the junction tree algorithm) that is able to perform inference in linear time with the network size, once a special graphical equivalent to the Bayesian network, called junction tree, is constructed. Unfortunately, this "nice" computational property is lost, when we extend the general case of discrete Bayesian networks to the general case of hybrid Bayesian networks, incorporating both discrete and continuous variables. Nevertheless, following given topological and other limitations (instantiated continuous variables), we can use the discrete version of the junction tree algorithm, without modification, thus preserving its computational efficiency. We have also described algorithms needed for BN CPD learning with observed and unobserved variables.

Finally, a particular extension of the Bayesian networks, i.e. decision network is presented that allows incorporation of preferences in the process of deciding on a given variable state in the network. Decision networks utilize the principle of maximum expected utility to model decisions that are optimal not only in the case of uncertainty of the hidden state value, but also when the decision system is modelled as an agent that has its own preferences for each state value.

Part II

Error handling in human-robot speech-based interaction

On designing voice-enabled interface for an interactive tour-guide robot

5

This chapter presents the initial study and design methodology development for building basic voice-enabled interfaces adapted to the nature of the autonomous tour-guide robots, behavioral requirements of visitors and noisy environment of mass exhibition. In the study we analyze voice-enabled interactivity between tour-guide robots and their users with the aim of deriving tour-guide dialogue task requirements. The analytical approach is used in the development of a preliminary prototype of a voice-enabled interface on a real multimodal robotic platform - RoboX. The prototype is further investigated in a field experiment during RoboX's deployment at the Swiss National Exhibition Expo.02. The lessons learned during Expo.02 showed that not surprisingly, speech recognition and synthesis performance is of crucial importance for enabling the interactive conversation between visitors and tour-guide robots. The type of human-robot interaction in mass exhibition conditions is typically short-term and abounds with variety of uncertainties. These uncertainties are mainly due to visitor behavior and attitude towards the robot in the human-robot interaction, as well as to the unreliable speech recognition in noisy conditions. The above two factors motivate system-initiated dialogue management, where the key issue is the identification of the user goal to attend a particular exhibit presentation. Correct identification of the user goal when the user in answering the questions during the exhibit presentation is essential for keeping high level of user interest while conveying exhibit-specific information. Noisy speech and some behavior of the visitors to mass exhibitions can jeopardize the process of user goal identification based solely on speech recognition during human-robot interaction, and can easily cause communication failures.

In order to address the problem of the risk for communication failures we argue for need of combining speech with other available modality information in the recognition error handling techniques,

fitted to the tour-guiding dialogue task requirements.

5.1 Interactive tour-guide robots

Human-robot interfaces are of great importance for robots that are to interact with ordinary people. In the setting of exhibitions, where people typically do not spend extensive amounts of time with a single robot, two criteria are considered most important: ease of use, and the level of visitor interest in the interaction. The human-robot interfaces must be intuitive, so that untrained and non-technical visitors of the exhibition can operate the system without prior instruction. The level of interest is an important factor in capturing people's attention.

Natural spoken communication is the most user-friendly means of interacting with machines, and from the human standpoint spoken interactions are easier than others, given that the human is not required to learn additional interactions, but can rely on "natural" ways of communication (Huang et al., 2001).

In an exhibition environment, the tour-guide robot often interacts with individual visitors as well as crowds of people. In such conditions it is important that the tour-guide robot takes the initiative and appeals to the "intuitions" of visitors. Thus, a primary component of a successful tour-guide robot is the ability to be aware of the presence of people and to engage in a meaningful conversation in an appealing way.

The main components of human-robot voice enabled interfaces are: speech output (loudspeakers) and input (microphones), speech synthesis for speech output modality, speech recognition and understanding for speech input modality, dialogue management and usability factors related to how humans interact with tour-guide robots (Spiliotopoulos et al., 2001). These components function by recognizing words, interpreting them to obtain a meaning in terms of an application, performing some action based on the meaning of what was said, and providing an appropriate spoken feedback to the user. Whether such a system is successful depends on the difficulty of each of these four steps for the particular application, as well as the technical limitations of the system. Robustness is an important requirement for successful deployment of such a technology, in particular speech acquisition and speech recognition, in real-life applications. For example, automatic speech recognition systems have to be robust to various types of ambient noise and out-of-vocabulary words. Automatic speech synthesis should not only sound naturally but also be adapted to an adverse acoustical environment. Lack of robustness in any of these dimensions makes such systems unsuitable for real-life applications.

In this chapter we describe our efforts in designing a preliminary voice-enabled interface for the tour-guide robot RoboX (Prodanov et al., 2002; Drygajlo et al., 2003) (Figure 5.1). RoboX was developed at the Autonomous Systems Lab (ASL) in EPFL and served as a tour-guide robot during the Swiss National Exhibition Expo.02 (Jensen et al., 2002a,b; Siegwart et al., 2003). The Expo.02 offered a convenient opportunity for performing a real-life field study of a voice-enabled interface of tour-guide robot.

5.2 Design philosophy background

The first specificity during the Swiss National Exhibition Expo.02 was that the tour-guide robots to be deployed in the robotic exposition should be capable to interact with visitors using four official languages: French, German, Italian and English. They had to attract people's attention, to show them the way to the exhibits and to supply information about these exhibits. Studying other specificities of autonomous, mobile tour-guide robots led us to the following observations.

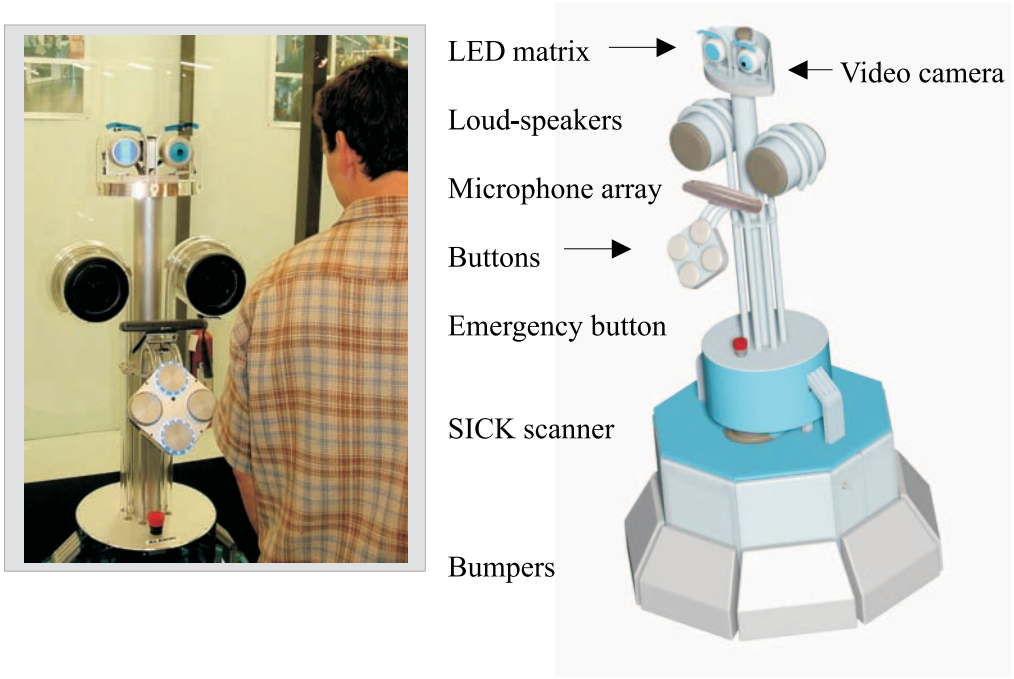


Figure 5.1: The mobile service robot RoboX.

First, even without voice enabled interfaces, tour-guide robots are very complex, involving several subsystems (e.g. navigation, people tracking using laser scanner, vision) that need to communicate efficiently in real time. This calls for speech interaction techniques that are easy to specify and maintain, and that lead to robust and fast speech processing.

Second, the tasks that most tour-guide robots are expected to perform typically require only a limited amount of information (Spiliotopoulos et al., 2001) from the visitors. Most of the time it is important that visitors acquire useful and interesting exhibit information. These points argue in favor of a very limited but meaningful speech recognition vocabulary and for a simple dialogue management approach. The solution adopted for Expo.02 was based on yes/no questions initiated by the robot where visitors' responses could be in the four official languages of the Expo.02 (oui/non, ja/nein, si/no, yes/no). This approach lets us simplify the voice enabled interface by eliminating the specific speech understanding module and allows only eight words as multi-lingual universal commands. The meaning of these commands depends on the context of the questions asked by the robot.

A third observation is that expo tour-guide robots have to operate in very noisy environments, where they need to interact with many casual persons (visitors). Figure 5.2 presents a typical example, where a clean speech (visitor's answer - Figure 5.2 a)) is corrupted by background noise of the exhibition room. It consists mostly of babble noise combined with noise resulting from the robots' movement and other sounds, such as beeps for example - Figure 5.2 b). This calls for speaker independent speech recognition and for robustness against noise.

The basic philosophy of the voice interface design methodology proposed in this chapter is to develop voice enabled interfaces that are adapted to the nature of autonomous, mobile tour-guide robots, behavioral requirements on the side of visitors and real-world noisy environments. The automatic speech recognition and synthesis systems have to cope with these factors.

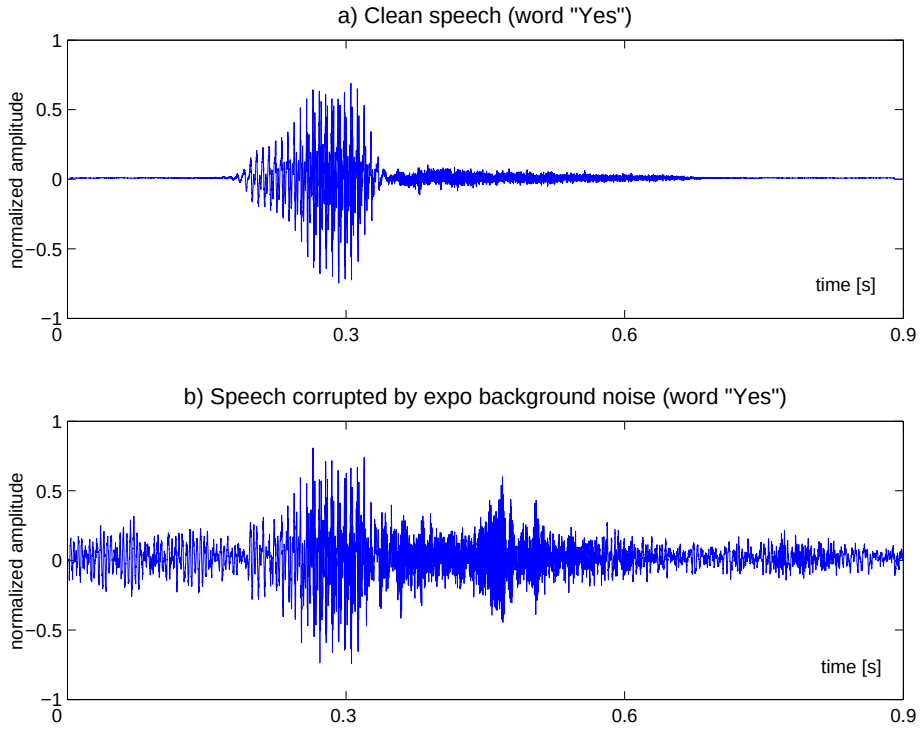


Figure 5.2: Word 'Yes' in (a) clean and (b) noisy conditions

5.3 Architectural overview

A block diagram of the functional architecture model for voice-enabled interface of RoboX is shown in Figure 5.3. It consists of speech output component (loudspeakers) and speech input component (microphones), speech synthesis for voice output, speech recognition for voice input and dialogue management that controls the sequence of verbal information exchange between the visitor and the robot utilizing speech and other modalities and, given a pre-defined sequence (task scenario) of events (scenario objects) (Jensen et al., 2002a).

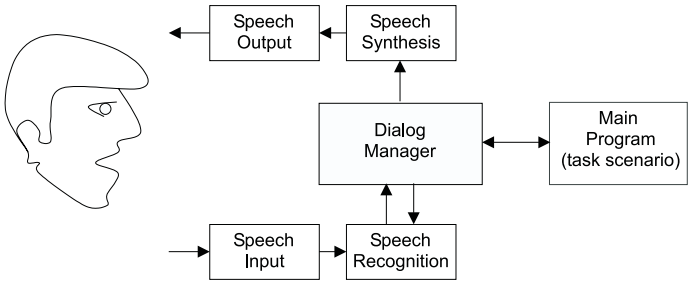


Figure 5.3: Voice-enabled interface

Speech is one of the input/output modalities within the multi-modal, multi-sensor interface of the robot and should naturally fit into the functional layers of the whole system. On the other hand, from a functional and conceptual point of view, the addition of a voice enabled interface does not affect the overall system organization; implementation should take some specific constraints into

account.

5.3.1 Hardware architecture

Figure 5.4 presents the hardware architecture of RoboX. It consists of three layers: input/output (I/O) layer and two (low- and high-level) processing layers.

Multiple sensors and other input/output devices of the I/O layer are used by the robot to communicate with the external world, in particular with users. In this set of multi-modalities, loud-speakers and a microphone array (Andrea Electronics DA-400 2.0) represent the output and input of the voice enabled interface. They are installed at half the height of the robot, which is a convenient position for both children and adults.

Among input devices that have to cooperate closely with this interface, when verifying the presence of visitors, are two SICK laser scanners mounted at knee height and one color camera placed in the left eye of the robot. The blinking buttons help in choosing one of the four languages, and the robot's face, which consists of two eyes and two eyebrows can make the speech of the robot more expressive and comprehensive. Finally, a LED matrix display in the right eye of robot may suggest the "right" moment to answer to the robot's questions (Jensen et al., 2002a)] (Figure 5.1).

The low-level processing layer contains hardware modules responsible for pre-processing of signals dedicated to input and output devices. The voice pre-processing is represented in this layer by the digital signal processor of the microphone array and the audio amplifier for the loud-speakers.

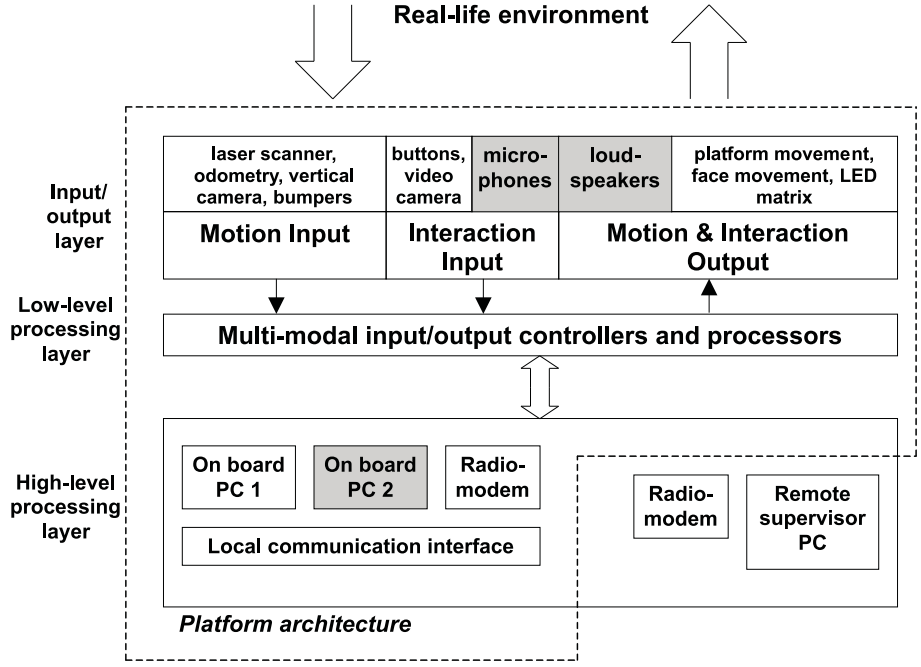


Figure 5.4: Hardware architecture

The high-level processing layer consists of two on-board computers: Pentium III (700M Hz, 128MB, 30GB HDD, Windows 2000) dedicated for all interaction tasks, including speech synthesis, speech recognition and dialogue management, and PowerPC 750 (400 MHz) for navigation.

Both computers can communicate with each other via local Ethernet and with external monitoring computer via wireless modems.

5.3.2 Software architecture

One of the most popular software architectures used in robotics is the three-layer architecture, which consist of a reactive, executive and deliberate layers (Russell and Norvig, 2003). The reactive layer provides low-level control by routines for the robot. It is characterized with a tight sensor-action loop. The executive layer serves as an interface between the reactive and the deliberative layer. It accepts directives from the deliberative layer and translates them into the needed sequence of reactive routines. The deliberate layer generates global solutions for high level complex robotic tasks.

In the case of RoboX, the principal robot operations are controlled by one main program called sequencer, which executes a predefined sequence (task scenario) of events (scenario objects). The overall architecture of the sequencer including speech synthesis and recognition objects and dialogue sequence management is depicted in Figure 5.5. The sequencer program is implemented in SOUL (Scenario Object Utility Language) designed at ASL to meet the requirements of the autonomous, interactive, mobile tour-guide robot. The main program is defined as a graph like scenario where the execution of the sequence of events corresponding to a predefined task is strictly linear (Jensen et al., 2002a,b). The events generated by the sequencer should be treated as logical events. Therefore, each of the scenario objects have a finite number of possible outcomes, which reflect the different states of this object after its execution. For example, the speech recognition object has three possible outcomes, corresponding to yes and no answers, and maximal execution time flag (time-out). Several scenario objects may be running in parallel, e.g. speech synthesis and face movement objects.

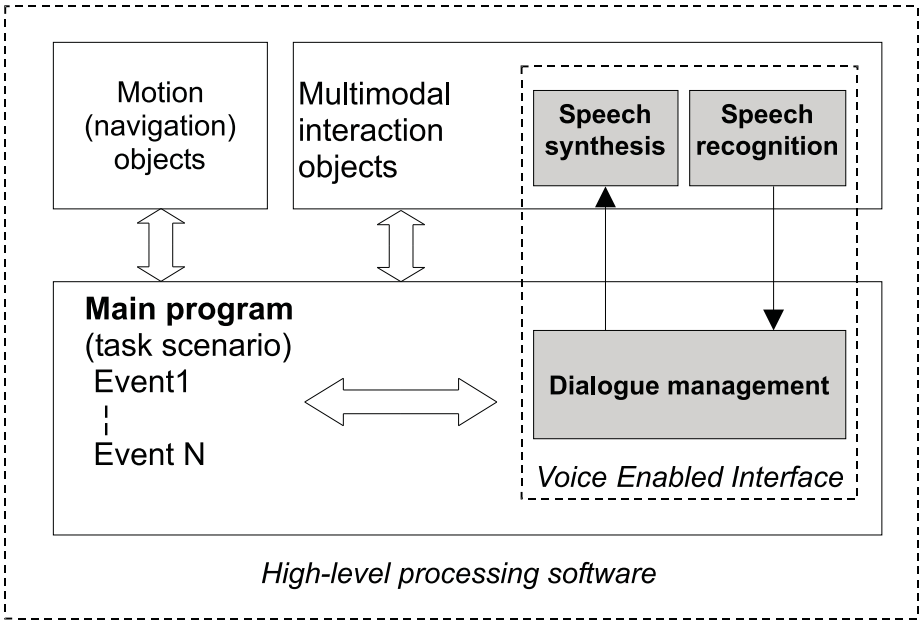


Figure 5.5: Software architecture

The scenario sequence as a whole can be associated with the tasks related to the deliberative

layer in the three layer model. Each scenario object itself forms a part from the executive layer and it is responsible for sequencing the low-level reactive tasks that as atomic action units form the behavior of the robot at each moment in time.

5.3.3 Tour-guide task scenario

The main task scenario of RoboX is to guide the visitors of the exhibition in accordance with the predefined tour plan and visitor's expectations, when coordinating the various robot activities related to sensing, motion and visitor-robot interaction. A dialogue scenario has to fulfill these required properties of the main task scenario by appropriate verbal expressions, explanations and questions of the robot and the visitors' confirmations.

The main requirement for a tour-guide dialogue scenario from the side of the robot is to provide as much as possible exhibit information to visitors in a limited time. The fulfilment of this requirement is dependent on the level of interest of the user (visitor) in the currently presented exhibit. Therefore, it is essential to provide short description of what the visitor can expect or what exhibit will be presented. Then, each exhibit presentation after a short description can start with a question concerning the user goal, i.e. a question to elicit the intention of the user to attend to the proposed presentation.

Therefore, one of the tour-guide dialogue tasks is to infer the goal of the user at the beginning of each exhibit presentation (e.g. as in Figure 5.6 row 2.a)). The second task is related to the exhibit presentation that follows. In order to maintain high level of user interest, we assume that h/she has to be involved frequently in the conversation. Hence, the second dialogue task of the tour-guide robot is related to providing a system-driven dialogue of conversational type. In such a task the tour-guide robot can ask questions to the user about the presented exhibits to keep him involved and interested (e.g. as in Figure 5.6 rows 2, b) and c)).

Given that the information presented to visitors is new to them, the structure of the dialogue can be defined as a sequence of the above two dialogue tasks: inference of the user goal and exhibit presentation. Thus the dialogue as whole can be well structured and a state-based dialogue management can be used for controlling the dialogue flow. Following the proposed dialogue structure the dialogue scenario can be designed to allow the presentation of a limited number of exhibits according to the visitor flow and resulting tour time limit.

In the main program, the tour-guide dialogue scenario in the form of the sequence of sub-dialogues named Introduction, Exhibit 1, Exhibit 2, ... , Next Guide, is embedded in the task scenario (Figure 5.6). Some examples of dialogue sequences are presented in section 5.5. Concepts

1. ***Introduction*** (My name is Robox and today I will be your tour-guide…)
2. ***Exhibit 1***
   a. *Description:* (These little things here are called Alice.
      Do you want to know about them?)
   b. Do you know that they can be radio controlled?
   c. Do you want to play with the little Alices? …
3. ***Exhibit 2***
4. ...
5. ***Next guide***

Figure 5.6: Dialogue scenario

of speech synthesis and speech recognition objects and the corresponding programs are presented in Section 5.4.

5.4 Voice-enabled interface

To start interacting with people, a method for detecting them is needed. We have found that in the noisy and dynamically changing conditions of the robotics exposition, a technique based on motion tracking using laser scanners, and on face detection with a color video camera gives the best results (Chapter 8). When RoboX finds people in the distance smaller than 1.5 meters, it should greet people and inform them of its intentions. The most natural and appealing way to do this is by speaking. In the context of the national exhibition (four official languages) and having the possibility of rapid prototyping of complex interaction scenarios when using the voice enabled interface, speech becomes one of the most important output modalities to be used for communicating with visitors.

5.4.1 Speech synthesis

In the noisy environment of the exposition, the automatic speech synthesis system should generate speech signals that are highly intelligible and of an easily recognizable style; if possible, this style should correspond to the style of an excellent human guide. On the other hand, and to preserve the robot's specificity, the quality of its speech should not mimic perfectly the human speech, but such speech has to sound natural. Two main criteria that we have used to choose an appropriate method for automatic speech synthesis were intelligibility and naturalness.

Therefore, a solution adopted for the speech synthesis event is a text-to-speech (TTS) system based on concatenation of diphones (phonetic units that begin in the middle of the stable state of a phoneme and end in the middle of the following one) (Dutoit, 1997). The actual task of the synthesizer is to produce, in real time, an adequate sequence of concatenated segments, extracted from its parametric segment database and the prosodic parameters of pitch pattern and segmental duration adjusted from their stored values, to the one imposed by the natural language processing (NLP) module. The intelligibility and naturalness of the synthesized speech highly depends on the quality of the segment database, grapheme-to-phoneme-translation and a prosodic driver for pitch and duration modification.

During the experimentation phase with RoboX, the best results, e.g. for French, were achieved for the combination of LAIPTTS (NLP) (Keller and Werner, 1997), Mbrola reproduction tools and a Mbrola parametric segment database. For all four application languages (French, German, Italian and English) the structure of the speech synthesis system is the same, and the system can be limited to Mbrola phonetic files generated off-line by the NLP module, Mbrola synthesis engine and parametric segment databases for different languages.

When RoboX needs a yes/no response from the visitor, the speech synthesis event is directly followed by the speech recognition event in the task scenario.

5.4.2 Speech recognition

The first task of the speech recognition event is the acquisition of the useful part of the speech signal that avoids unnecessary overload for the recognition system. The adoption of limited in time (2 seconds) acquisition is motivated by the average length of yes/no answers.

Ambient noise in the exhibition room is one of the main reasons for degradation of speech recognition performance. To add robustness against ambient noise without additional computational overhead a microphone array (Andrea Electronics DA-400 2.0) is used. During the 2 seconds acquisition time the original acoustic signal is processed by the microphone array. The mobility of the tour-guide robot is very useful for this task since the robot, when using the people tracking system, can position his front in the direction of the closest visitor and this way can direct the microphone

array. The pre-processing of signals of the array includes spatial filtering, de-reverberation and noise cancelling. This pre-processing does not eliminate all the noise and out-of-vocabulary (other than yes/no) words. It provides sufficient quality and non-excessive quantity of data for further processing.

Recognition should be speaker independent and multi-lingual performing equally well on native speakers and on speakers who are not native of the target language. The system is intended to recognize the limited vocabulary of eight words (oui, non, ...) but can accept an unlimited vocabulary input. In such a system, we are not only interested in a low error rate, but also in rejection of irrelevant words.

At the heart of automatic speech recognition system of the robot lies a set of the state-of-the-art algorithms for training statistical models of words and then using them for the recognition task (Renevey and Drygajlo, 1997). In a speech recognition event the signal from the microphone array is processed using a Continuous Density Hidden Markov Model (CDHMM) technique where feature extraction and recognition using the Viterbi algorithm are adapted to a real-time execution. The approach selected to model eight key words (oui, non, ja, nein, si, no, yes, no) is the speaker independent flexible vocabulary approach. It offers the potential to build word models for any speaker using one of the four official languages of Expo.02 and for any vocabulary from a single set of trained phonetic sub-word units. The major problem of a phonetic-based approach is the need for a large database to train, initially, a set of speaker-independent and vocabulary independent phoneme models. This problem was solved using standard European and American databases available from our speech processing laboratory, as well as specific databases with the eight key-words as recorded during experiments. The CDHMM toolkit (HTK) (Young et al., 2002) based on the Baum-Welch algorithm was used for the training.

Out-of-vocabulary words and spontaneous speech phenomena like breath, coughs and all other sounds that could cause a wrong interpretation of visitor's input have also to be detected and excluded. For this reason a word spotting algorithm with garbage models have been added to the recognition system. These garbage models were built from the same set of phoneme based sub-word models (Huang et al., 2001; Wilpon et al., 1990; de Mori, 1998) thus avoiding additional training phase or software code modification. Finally, the basic version of the system is capable to recognize yes/no words in four languages and speech acoustic segments (undefined speech input) associated to the garbage models.

A detailed description of the recognition system of RoboX is presented in Appendix B.

5.5 Dialogue management

Speech synthesis and recognition alone are not sufficient for realizing the dialogue scenario as presented in Figure 5.6. Similarly to humans, the expo robot needs a dialogue control system for maintaining the spoken interaction on a multi-modal platform. This system called dialogue manager is responsible for handling and maintaining the short-term sequences of scenario events like speech recognition, eye movement, LED matrix animation, people tracking, speech synthesis, etc. in order to succeed in the main goal of tour-guiding as presented in Figure 5.6. SOUL language allows for creating different sub-scenarios associated to these short-term tasks and embedding them in one main sequence. This results in fixed state-based dialogue management, meaning that all the sequences should be scripted in advance. In order to represent the functional structure in the sequences of tour-guide dialogues, we have adapted graphical state-based formalism similar to flow-charts. Some possible sequences are presented in Figures 5.7-5.9. They include not only speech events but also some non-speech events, e.g. move event, motion tracking event, behavior event,

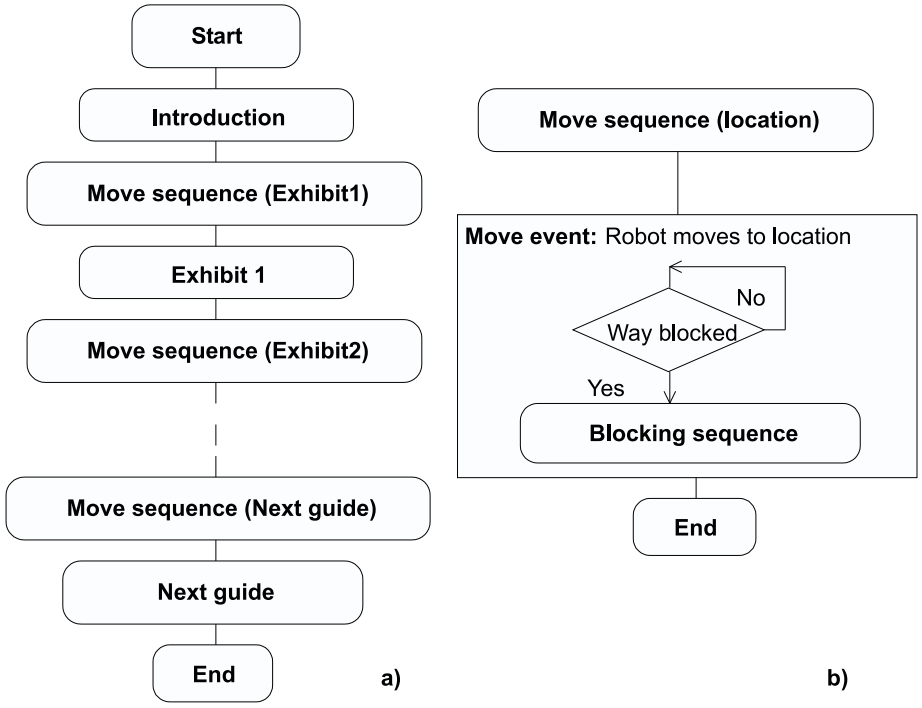


Figure 5.7: (a) Main sequence. (b) Move sequence

etc.

The major advantage of the state-based dialogue management is in its simple implementation. From the point of view of dialogue development, state-based structures such as the ones presented in Figures 5.7-5.9 are particularly suitable for a dialogue flow with well-structured dialogue tasks involving predefined sequences of exchange of information between the user and the dialogue system (McTear, 2002). Given the structure of the tour-guiding dialogue scenario in Figure 5.6 with the system retaining control on which question to ask next, state-based dialogue solution becomes an attractive solution. Moreover, the state-based dialogue control restricts the user input to predefined words or phrases matching carefully designed system prompts. Such a strategy allows for a speech recognition process that requires simple technical solutions and relaxed computational demands. In the short-term interaction between tour-guide robots and visitors unfamiliar with robots, in noisy exhibition conditions, the state-based dialogue offers a fair trade-off between the mentioned advantages and the lack of certain flexibility and naturalness of interaction.

5.6 The Expo.02 experiments

During the five-month period from May 15 to October 15, 2002, eleven RoboX systems were interacting with the visitors of Expo.02. Two of them were equipped with microphone arrays and a full version (speech synthesis and recognition) of the voice enabled interface, described above. An important aspect of the tour-guide robot voice-enabled interface is the robot's physical interaction with visitors. During the Expo.02 period, we conducted experiments with different scenarios and different versions of the interface software and monitored the performance of the voice-enabled interface in adverse environment conditions. Finally, ten tour-guide robots were capable of successfully

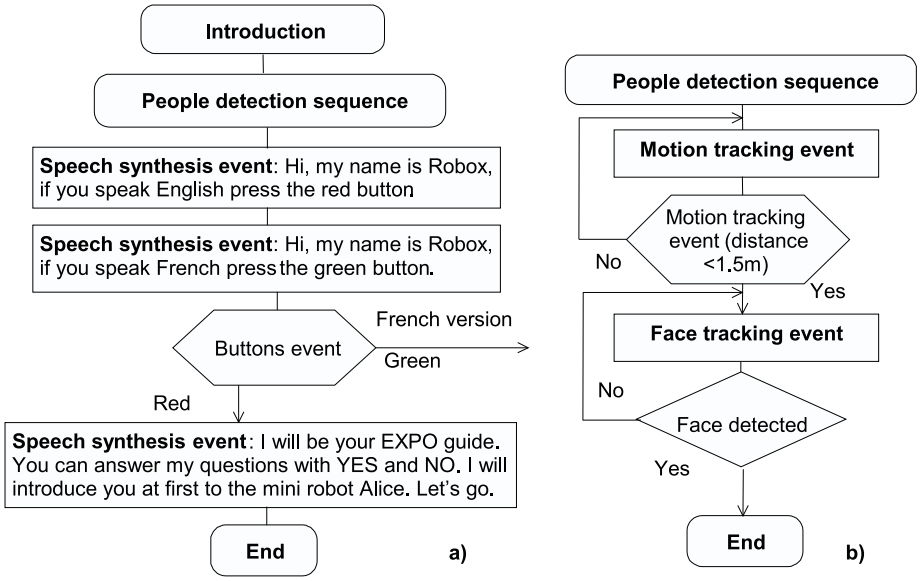


Figure 5.8: (a) Introduction sequence. (b) People detection sequence

presenting five out of ten exhibits during a single tour. This corresponds to five different exhibit sequences chosen between ten available scenarios for each tour. The exhibit to be presented by one robot is chosen if there is no other tour-guide robot presenting this exhibit (Jensen et al., 2002b). Each exhibit sequence begins with yes/no question of the robot asking the visitors if they want to see the exhibit or not. Then the average number of presented exhibits per tour depends on the number of yes answers, recognized at the beginning of each exhibit sequence.

A database including visitors' responses and information related to the recognized words and the scenario events in the particular dialogue sequence was recorded on the robot's interaction PC and then transferred to the remote supervisor PC (Figure 5.4). These data were used for optimizing the speech recognition system by noise-matched re-training of the HMMs. The database was also used for assessing and modifying the existing dialogue sequences. After Expo.02, this database was used for our further research (Chapter 6).

5.6.1 Expo.02 observations and statistics

Robot-visitor interaction, with many visitors and several robots in a public exhibition is a complex task. When RoboX is giving a tour it stops at several places and supplies information related to a certain part of the exposition (exhibit). With several RoboXs running at the same time we faced the problems of multi-robot coordination, visitors flow, visitors density and visitors behavior. Expo.02 is considered a mass exhibition with several thousands of visitors per day. During the preparation of the project we anticipated up to 500 visitors per hour, which results in 125 visitors enjoying the robots at the same time, assuming a 15 minutes stay inside the Robotics Exposition. In the period from 15.05.02 to 15.10.02 an average number of about 4500 people visited the exposition every day. This results in a visitor flow of 450 persons per hour on 315 m^2 exposition space, with up to ten operating robots. In such conditions, the autonomous robot's ability to interact with people via spoken dialogue, in addition to direct physical interaction, was, for most visitors, the most fascinating aspect of the entire exhibition. We have learned several lessons from such robotic dialogue design.

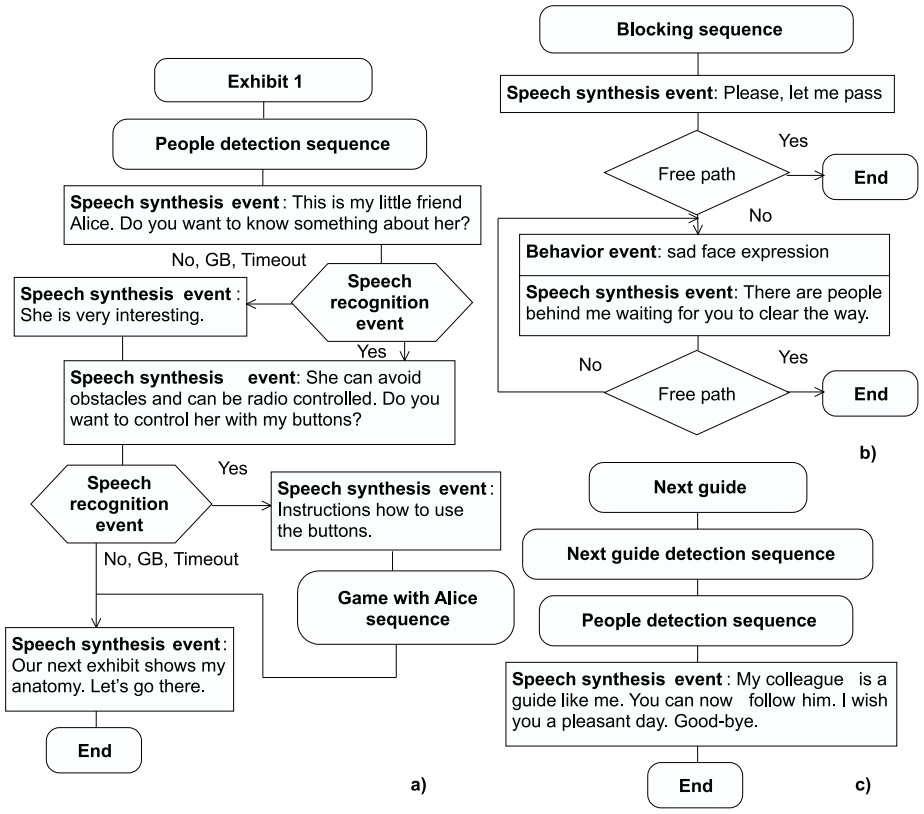


Figure 5.9: (a) Exhibit 1 sequence. (b) Blocking sequence. (c)Next Guide sequence

Visitors behavior

Visitors' behavior varies from collaborative to investigative and to even "destructive", and can hardly be anticipated. The majority of visitors treat a robot sometimes as human and sometimes as a machine, clearly following some modified form of human interaction. Often they treat the robot like a human by default, dodging its path and verbally responding to it, using vocabulary proposed by the robot. If they become interested in some features of the robot, or want to investigate how it works, however, they begin to treat the robot like a machine or a toy, by repeated standing rudely on its path or by pressing the emergency button to force a reaction. The ability of RoboX to decelerate and stop in the presence of people and to "ask" for clearance proved to be one of the most entertaining aspects of the entire system. Many visitors were amazed by the fact that the robot acknowledged their presence by using speech and an alarm sound, and for this reason they repeatedly stepped in its way.

In general, visitors appreciated the two robots with a full voice enabled interface (speech synthesis and recognition). They were willing to interact using speech, even when it was more difficult in the noisy expo environment than pressing the buttons. They tried many times to respond yes or no, or even to use both answers at very short time when the performance of the recognition was not satisfactory, before leaving the tour-guide robot.

Children' behavior and reactions were particularly interesting. They were the most emotional, curious and "destructive" users. In most cases, they were not really listening to the instructions of RoboX. They tried to interact with the robot on a more basic level by pressing buttons, waving

hands to see the reaction of its eyes, catching him while moving, or even climbing on the robot, calling for their parents to come and watch. Once they learned about speech input, they were using it often in groups of kids or with their parents.

Robot's behavior

It is important that the robot's scenario should be adapted to the guiding task and to the visitors' behavior described above. Firstly, the robot must sense the presence of visitors and stop its presentation if they go away. RoboX uses several sensors and algorithms to achieve awareness of its environment. Simple switches detect events like visitors pressing the emergency button, the interactive buttons or hitting the bumpers. The obstacle avoidance provides sufficient information when visitors are blocking the robot. In addition, the robot is aware of visitor presence in its surroundings by means of face and motion tracking.

Second, long repetitive presentations are guaranteed to drive visitors away. Instead, short "game like" scenarios, such as the one presented in Figure 5.9 a) with interactive questions, are most effective and proved to be more interesting and appealing. If the visitor is often engaged in answering the yes/no questions, and the guide is responding promptly with variable behavior, combining speech, movement, facial expression, and strange sounds, then the interaction is interesting and occupies people for a much longer time. In addition, every scenario should be equipped with dialogue sequences for escaping emergency situations such as the robot's path being blocked (see Figure 5.9 b)), or excessive playing with its interactive buttons or bumpers.

Third, there is often a crowd of people around the robot rather than a single person. Together with expo background noise, this makes it difficult or impossible for some people to hear the robot's questions if they are purely verbal. Therefore, all robot interactive questions should be multimodal. Normally they are accompanied by the LED matrix animation as well as face (eye and eyebrows) movement (Figure 5.1).

In every scenario speech synthesis is the most important event. While speech recognition can be compensated using the interactive buttons, tour-guiding is impossible without the ability of the robot to speak. Both speech synthesis and recognition should be adapted to the noise in the exhibition room as people easily get frustrated if they cannot hear the tour-guide robot very well, or if their answers are often wrongly recognized.

Voice enabled interface scenario evolution

The underlying goals of compelling interaction and maximal autonomy of the expo tour-guide robots have remained constant throughout the creation of all dialogue scenarios. However, each succeeding scenario was the product of a complete re-design phase based on lessons learned from prior scenarios. The typical example is the "introduction scenario" where a visitor has to choose one of the four official expo languages. In the first scenario, RoboX asked four questions, "Do you speak English/German/French/Italian?" in the four official languages. Although these questions implied a yes/no answer, people were often expecting the robot to better understand utterances such as "No Italiano" or "Ich spreche Deutsch". To avoid this we added help information before the questions: "For English/French/German/Italian answer with yes/oui/ja/si or no/non/nein/no" in the four languages supported by the interface. This made the "introduction sequence" longer than before, but more effective. During Expo.02 all of the scenarios were evolved in order to improve the efficiency of interaction of each tour-guide robot. This was achieved in spite of the need for laborious editing of the dialogue sequences using the SOUL language.

Voice enabled interface problems

The main difficulty in the voice-enabled interaction with visitors came from the recognition errors due to noisy conditions. This was the case when garbage models were most likely or the answer was too late, although the LED matrix display was used to indicate the right moments for answering. The recognition system was additionally trained with noisy speech from that recorded for the Expo.02 database. This resulted in improved recognition and overall performance, resulting in more visited exhibits per tour. However, the worst cases resulted in e.g. false language selection due to recognizing yes or no, when the visitor was silent and only background noise was captured. In those cases, users normally ceased interaction with the robot. Uncooperative visitors often challenged the interaction by not attending properly to the conversation and remaining silent when they were supposed to answer. This was also the case of initially cooperative users that were leaving in a middle of a conversation or were responding to other people talking with them during the robot questions. Such behaviors were resulting in communication failures such as wrong language and exhibit selection as well as continuing a presentation without a real user in front of the robot.

From these observations it became apparent that in order to successfully convey its information a tour-guide robot needs to employ special error handling techniques dedicated to avoid communication failures due to recognition errors in noisy conditions and non-cooperative users.

User survey

To get the feedback on the scenario performance, apart from the observation studies concerning human-robot interaction, questionnaire forms collected from visitors were analyzed with respect to the outlook of the tour-guide robot, its behavior, its speech synthesis and recognition quality, the number of exhibits visited and how people found the tour-guide presentation. During five typical expo days 209 people were questioned, 83 of them had interacted with robots using speech recognition. The results of this evaluation are presented in Table 5.1 and Figure 5.10. They are very useful in that they establish a quantitative account of the scenario efficiency for the Expo.02 tour-guide robots.

Total people	Langue*			average	Gender**		average number of
209	DE	FR	IT	Age	M	F	visited locations
count	128	75	6	34.4	88	105	per visitor
%	61%	36%	3%	-	42%	50%	4
*DE - German, FR - French, IT -Italian; **M - male, F -female							

Table 5.1: Survey population statistics

5.7 Summary

The preliminary study and design methodology proposed in this chapter is conceived for developing voice enabled interfaces that are adapted to the nature of autonomous, mobile tour-guide robots with all their constraints, behavioral requirements of visitors and real-world noisy environments that the automatic speech recognition and synthesis systems have to cope with. In the approach presented, the development was focused on the potential user, from the very beginning of the design process through to the complete system. As a result the analytical development and field experiment

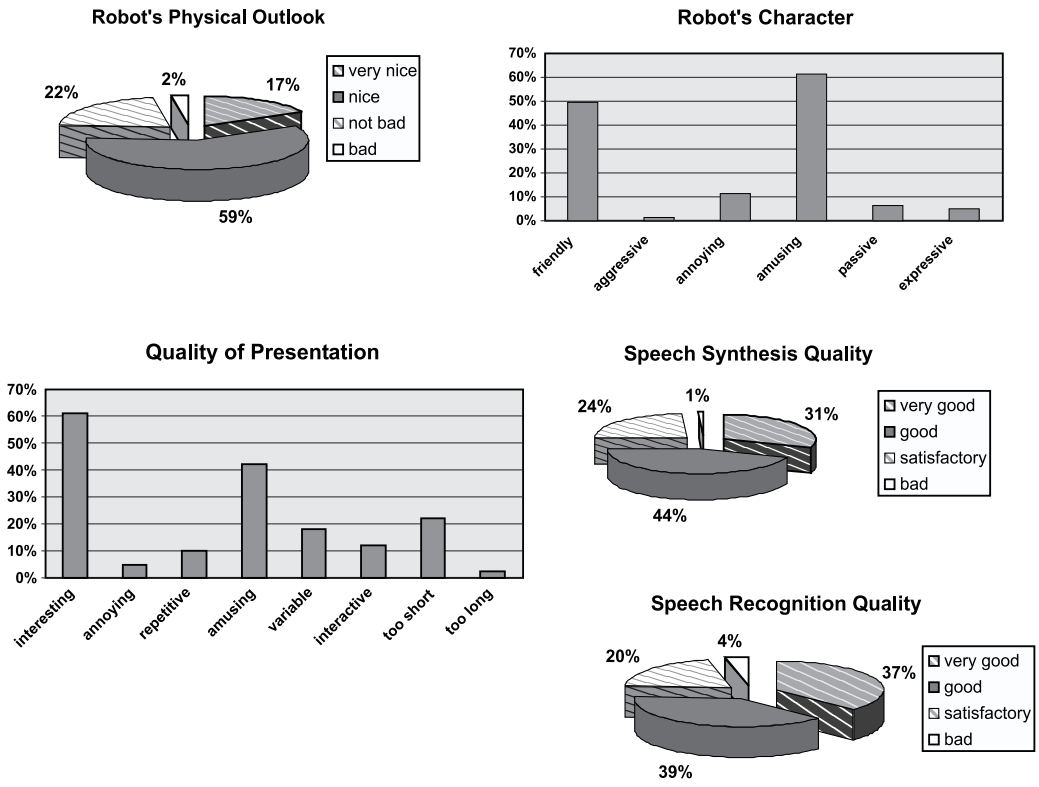


Figure 5.10: Results of the visitor survey

presented in this chapter has validated the initial assumptions on which designing intelligent voice enabled interfaces for tour guide robots can be based.

The main requirement for a tour-guide dialogue system is to present as much as possible exhibit information to the interacting visitor (user) in a limited time. The short-term human-robot interaction in mass exhibition conditions, where visitors and robots produce high level of acoustic noise motivate a robot-driven dialogue flow, relying on keywords or short meaningful phrases (e.g. yes/no keywords) as universal user commands to the robot. The tour-guide dialogue can be constructed out of two subsequent dialogue tasks and related dialogue states: (1) user goal identification before each exhibit presentation; (2) interactive exhibit presentation in the form of informal conversation. Such well-structured dialogue task domain additionally motivates the use of state-based dialogue management employing dialogue turns based on question/answer pairs (e.g. robot question/visitor answer) initiated by the robot.

The field experiments during Expo.02 showed that such an interaction scheme could be seriously challenged by the visitors' behaviors. Since the visitors' behavior during the dialogue can vary, there are often cases when people do not follow the choice suggested by the robot, using out of vocabulary words and even giving both yes and no answers or simply remain silent. The presence of crowds of people and moving robots in the exhibition room results in adverse acoustic conditions, causing errors in the speech recognition. Hence, a system managing speech-based interaction with visitors should employ error-handling techniques in order to avoid communication failures.

Standard techniques for error handling in speech recognition are based on error detection and correction and usually use recovery dialogues (McTear, 2002). Detecting errors using only speech recognition can be difficult and repair dialogues may be inefficient in the acoustic conditions of mass

exhibition. Speech recognition error handling methods adapted to the tour-guiding requirements and combining speech and other available modality information can be beneficial in these conditions.

Modality fusion for error handling in communication with tour-guide robots

6

After having presented the design methodology for voice-enable interfaces of tour-guide robots (Chapter 5), this chapter develops methods of modality fusion for error handling in communication with tour-guide robot. Under the assumption of short-term interaction with visitors in adverse audio conditions an identification of the user's (visitor) goal at each dialogue state can be improved by combining interpretation of speech recognition results with information from other available modalities.

In this chapter, we introduce a probabilistic model for recognition error handling in human-robot spoken dialogue under adverse audio conditions. In this model, a Bayesian network framework is used for the interpretation of multimodal signals in the spoken dialogue between a tour-guide robot and the visitors in the mass exhibition conditions. In particular, we present methods for combination of the speech and laser scanner input modalities in the dialogue management system of the autonomous tour-guide robot RoboX. To infer the visitors' goal under the uncertainty intrinsic to these two modalities, we introduce Bayesian networks for combining noisy speech recognition results with data from a laser scanner, which are independent of acoustic noise. Experiments with real-world data , collected during the operation of RoboX at Expo.02, demonstrate the effectiveness of the approach in adverse environment. The proposed method makes it possible to model the error handling processes in spoken dialogue systems, which include complex combination of different multimodal information sources.

6.1 Error handling in the human-robot dialogue

Standard techniques for error handling in spoken dialogue are based on error detection and correction using recovery dialogues (McTear, 2002, 2004; Bulyko et al., 2005; Sturm and Boves, 2005). Detecting errors using only speech recognition can be difficult and repair dialogues may be inefficient in the acoustic conditions of mass exhibition. The type of interaction faced by a tour-guide robot in the exhibition room is usually short-term as visitors have limited time and want to see as many exhibits as possible. They do not have any prior knowledge in robotics and typically the initiative in the dialogue is left to the robot then. In such conditions people do not tolerate repetitive, time-consuming repairs that may often occur in the noisy conditions. Such repairs would most probably drive the visitors away. The least delay in the communication will be associated with an alternative method that can detect and correct errors immediately. In this chapter we introduce such a method, based on Bayesian networks and error handling through multimodal signal fusion, using auxiliary information from acoustics-insensitive signals to compensate for speech recognition errors (Drygajlo et al., 2003).

The chapter is structured as follows: In Section 6.2 the tour-guide dialogue is revisited, focusing on inferring the intention of the visitor (user goal) using speech recognition in noisy conditions. We argue in Section 6.3 that the speech recognition errors arising in such a dialogue can be handled by using auxiliary information from a laser scanner signal. In Section 6.4 Bayesian networks are introduced as a probabilistic framework for fusing the above two modalities in inferring the goal of the visitor. The approach is tested through experiments with real data, collected during the deployment of the tour-guide robot RoboX at the Swiss National Exhibition Expo.02 (Jensen et al., 2002a). Finally the potential benefits of multimodality fusion for error handling in spoken dialogues with robots are outlined with future perspectives in the Discussion and Summary sections of the chapter.

6.2 Tour-guide dialogue structure

The tour-guide dialogue can be seen as composed of dialogue states, where each dialogue state executes a sequence of events (i.e. scenario objects, Chapter 5), such as a speech synthesis event, a speech recognition event, a robot movement event, etc (Figure 6.2). The sequence of events forming a dialogue state is dedicated to the presentation of a specific exhibit. The number of these dialogue states is fixed. It can be defined in advance based on the number of exhibits described by the particular exhibition plan. Each dialogue state contains a dialogue exchange in the form of initiative/response (robot's question/visitor's answer) pair, during which the speech recognition is typically used to infer the "goal" of the speaker in the context of the current state (Figure 6.1 (a)).

We assume that the spoken responses (utterances) coming from visitors during the interaction can be mapped onto a finite number of state dependent user goals, which are used to infer the next dialogue state. In Figure 6.1 (b) this process is depicted graphically; UG stands for the user goal and DS for the dialogue state. We assume that the state of the dialogue at time t depends on the dialogue state and the user goal at time $t-1$, and it can also affect the current user goal at time t. Then the key issue in spoken dialogue management is to decide on the most likely user goal in the current dialogue state.

The initiative/response pair (IRP) in the case of RoboX during Expo.02 consisted of yes/no question from the robot and the answer from the visitor. An exhibit presentation was initiated by a short description followed by a question (the tour-guide robot asks the visitors if they want to see the next exhibit) and the visitor answer. Depending on the user goal the robot was either

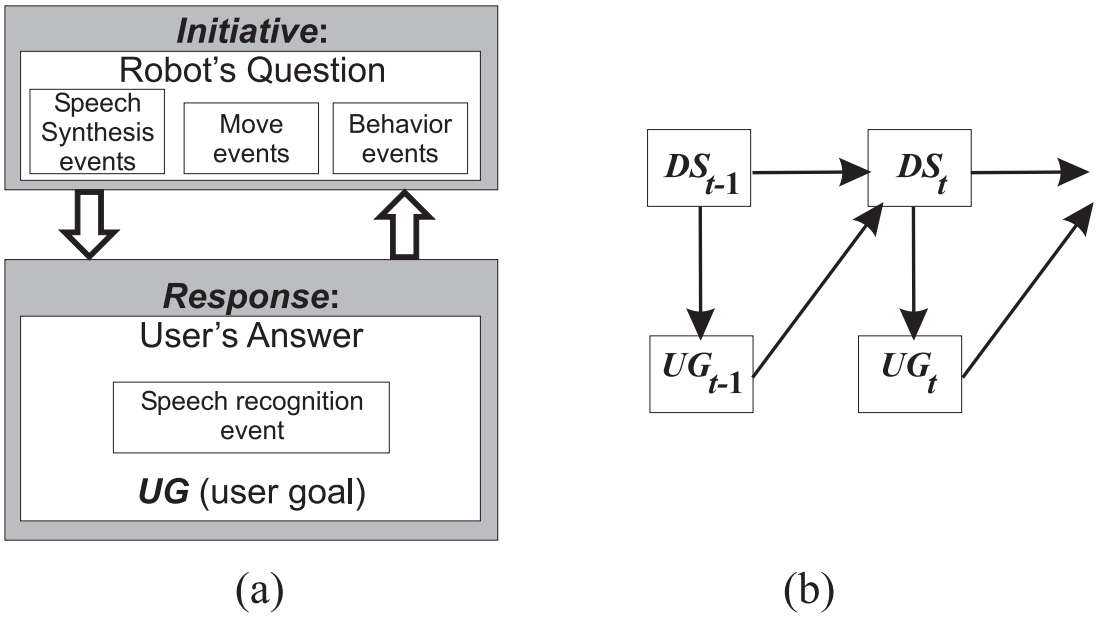


Figure 6.1: (a) Initiative/Response pair and (b) Dependency graph for spoken interaction management

proposing another exhibit in the next IRP or was initiating an interactive dialogue sequence for presenting the current exhibit. There were also initiative/responce pairs (IRPs) in the interactive part of the exhibit presentation (Figure 5.9 (a)). The initiative/response pair can be seen as an atomic building-block for constructing dialogue sequences for exhibit presentations. At Expo.02 one complete tour consisted of five presentations (Jensen et al., 2002b). Successful speech recognition can be then measured by the average number of correctly recognized responses at the beginning and during each exhibit presentation.

The speech recognition system of RoboX had to distinguish between the keywords yes, no and out-of-vocabulary words, fillers, coughs, laughs and acoustic phenomena different from the keywords, generally called garbage words (GB). The Observed Recognition Result ORR={yes, no, GB} is mapped then into three possible user goals (UG), accounting for the visitor intention : "the user is willing to see the next exhibit" (ORR=yes then $UG = 1$); "the user is unwilling to see the next exhibit" (ORR=no then $UG = 2$) and "user goal is undefined" (ORR=GB then $UG = 0$). This ternary choice system can be extended to any multiple choice key-word spotting system. One such example, demonstrating speech-based language selection by the user in the introduction part of the tour-guide dialogue is shown in Figure 6.2. Since there were four official languages at Expo.02 (German, French, Italian and English) it is more natural to ask one question in a multi-choice fashion than using four yes/no questions.

In its present state the input speech modality of RoboX is extended to cover more keywords than just the yes/no pair. In general, the answer of the user can contain a keyword used as a command to request one of $N - 1$ possible services or can be undefined, corresponding to the garbage word (GB) (Figure 6.3). We have limited the number of possible goals per user turn to $N = 3$ in order to limit the possibility of identification errors. In this case, we define three possible user goals $UG = 1$ - first possible service, $UG = 2$ - second service and $UG = 0$ - undefined user goal at each dialogue turn. The concrete user goal and service definition depends on the particular context of the system-initiated questions during dialogue. The different meanings for the UG used in our experiments are

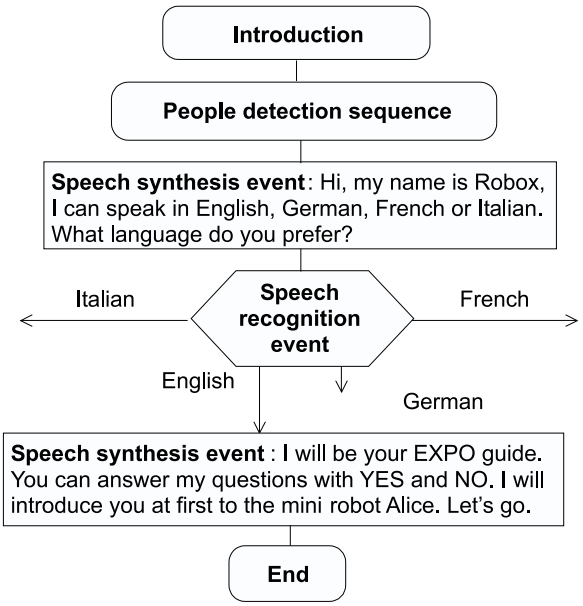


Figure 6.2: Dialogue example with more than two keywords

described in detail in Chapter 8.

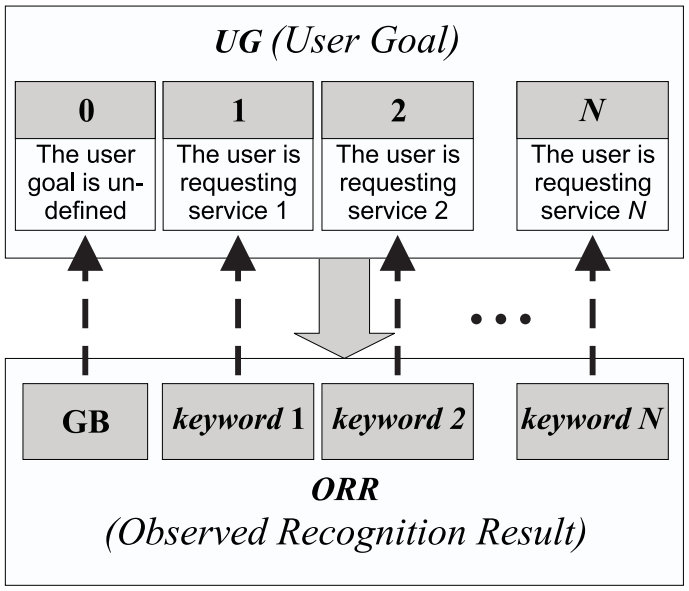


Figure 6.3: *ORR* to *UG* mapping

6.3 Multimodality fusion for speech recognition error handling

During human-robot interaction we need contextual information in order to interpret the recognition result into a correct user goal. The input speech modality provides the "acoustics-related" aspect of the user goal at the given dialogue state. The speech recognition result contains information about words possibly pronounced by people around the robot. However, it is known that recognition accuracy decreases in noisy acoustic conditions. Speech modality also does not provide reliable information on whether the words were pronounced by a human user or were part of the ambient acoustic environment.

On the other hand side, a mobile robotic platform usually have other input modalities that can contain auxiliary information, adding additional aspect through which the user goal can be inferred.

For example, at the Expo.02 there was often the case when initially interested visitors were leaving the robot, to respond to other people calling them. When this unexpected behavior was coinciding with the initiative/response pair, the GB word was often misrecognized for yes or no answer by the robot. In this case, in order to infer the right user goal ($UG = 0$), auxiliary information from the laser scanner signal revealing presence of visitors in close distance with respect to the robot (<1.5m), facing the microphone array is potentially beneficial.

The laser scanner signal carries information about the location of the communicating visitor. This "spatial" aspect of the user goal is essential, as absence of a user in given range in front of the robot could signal possible communication failure. Therefore fusing the different user goal aspects, supplied by the different input modalities on the robot platform, can result in more robust user goal identification, compared with using only one modality.

Information is extracted out of each modality in the form of events that can be inferred from the raw modality data. For example, the event that a user is staying in close range in front of the robot (UR user is in range for communication) can be inferred from the information contained in the laser scanner data. These modality events can be associated with the different user goal aspects. Hence, in a possible fusion method, we may need to account for three aspects of the user goal, e.g., speech recognition aspect, spatial aspect accounting for the presence of a communicating visitor and speech modality reliability aspect that gives information when the speech recognition result may be incorrect.

Figure 6.4 illustrates such a fusion scheme. The true underlying user goal with its three aspects can be seen as the cause of the observed values for the laser scanner signal and speech (audio) signal at each dialogue state. The spatial aspect about presence of people for spoken communication can be associated with the binary event UR "user in range for communication" (UR=1 user is in range, UR=0 user is out of range). Combining the observed recognition result (ORR) with evidence from the noise independent Laser Scanner Reading (LSR) affecting the event UR can change the confidence about the result of speech recognition. To define the influence of the acoustic environment on the speech recognition reliability we define the binary event SMR "Speech modality reliability", where $SMR = 0$ corresponds to unreliable speech, $SMR = 1$ when speech recognition is reliable. To infer the state of SMR the tour guide robot needs additional evidence about changes in the environment that can affect the reliability of the speech recognition, in particular the effect of acoustic noise on the speech signal. The Likelihood (Lik) as the measure of similarity between speech features and their models (speech recognition system score) along with an estimate of the speech-to-noise ratio (SNR) of the captured acoustic signal contain information about the environmental acoustic conditions (Huang et al., 2001). Likelihood measures are reported to be of limited efficiency in detecting speech recognition errors (Garcia-Mateo et al., 1999; Sturm et al., 2001). Therefore

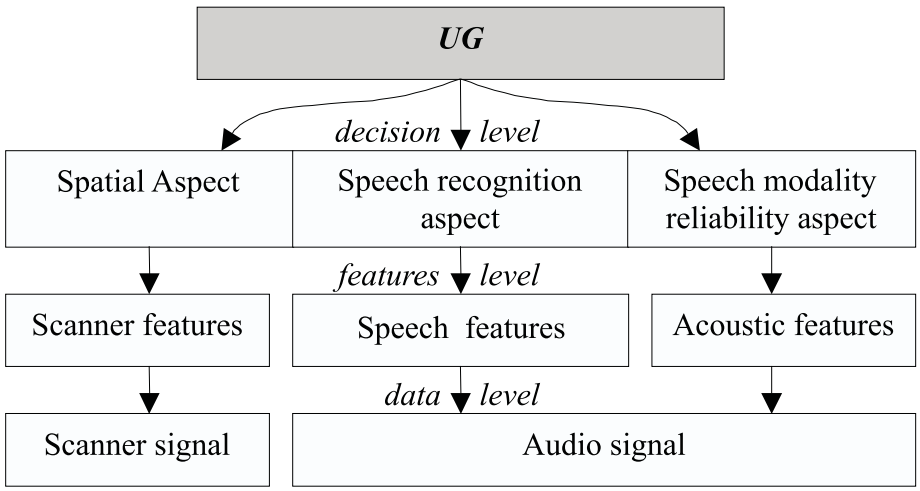


Figure 6.4: Fusion schema for user goal identification

combining Likelihood with additional measure of the level of acoustic noise such as the speech SNR is expected to provide more robust indication about the reliability of the recognition result than the likelihood only. The UR and SMR events can directly influence ORR (Figure 6.5).

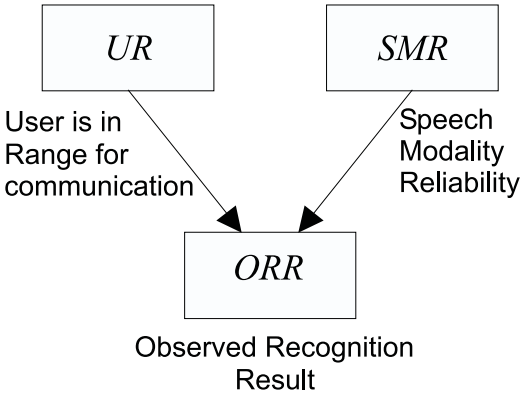


Figure 6.5: Bayesian network incorporating the causal dependencies between UR, SMR and ORR

Since this influence is not directly established, the causal relationships should be seen as probabilistic. It can happen that people are near the robot, answering as expected and recognition errors still occur. In that case, the influence of the cause can be modelled through a conditional distribution over the set of events. Bayesian networks have been shown to perform inference about probabilistically related events compatible with the notion of causal reasoning (Jensen, 1996). They have recently emerged as a promising tool for fusing multiple sources of information in dialogue modelling and pattern classification (Horvitz and Paek, 1999; Keizer et al., 2002; Nefian et al., 2002; Murphy, 2002). In the sections that follow a Bayesian network is used to probabilistically fuse data coming from the speech and laser scanner modalities of RoboX for inferring the user goals of visitors. The benefits for error handling in the human-robot dialogue using such an approach are

shown through experiments with real data, collected during the deployment of the tour-guide robot RoboX at the Swiss National Exhibition Expo.02.

6.3.1 Multimodality fusion: problem statement

Multimodal user goal identification can be seen as a classification problem, where at each dialogue state we want to assign a discrete variable UG (user goal) to one of M possible classes (user goals) $\{ug_1, ug_2, ..., ug_M\}$, given observed features originating from N different modalities $MF = \{MF_1, MF_2, ..., MF_N\}$.

In addition to this general problem, we might be interested in identifying N discrete-valued modality-related user goal aspects $ME = \{ME_1, ME_2, ..., ME_N\}$ given the observed modality features (MF), where $ME_1, ME_2, ..., ME_N$ denote modality-related events (e.g. all the variables in Figure 6.5). The modality features that are extracted from the underlying input modality signals may present substantial variability due to environmental and user factors. Hence, in identifying UG and ME, we have to work with distributions over the features. In other words, we have a multi-classification problem in which we want to calculate family of probabilities of the form $P(X_Q|E)$, where X_Q is a query variable of interest and E is the set of evidence variables containing the modality features and optionally some of the observed modality events (e.g. the observed recognition result, i.e. ORR in Figure 6.5).

For example, given the discussion in the beginning of this section X_Q can be equal to UG, while $ME = \{UR, SMR\}$ and $E = \{LSR, Lik, ORR, SNR\}$ and are interested in calculating $P(UG|LSR, Lik, ORR, SNR)$.

Given the above problem definition, Bayesian Networks (BNs) can be applied (Chapter 4) as a probabilistic framework for the modality fusion.

6.4 Bayesian networks for multimodal user goal identification

To build a Bayesian network model for inferring the most likely user goal (UG) value, we need first to define the set of random variables, the conditional independence assumptions between them and the variables of interest for inference: $P(X_Q|E)$.

In our case $X_Q = UG$. We define the Bayesian network variables' set as $V = (UG, UR, SMR, ORR, LSR, Lik, SNR)$. The meaning of these variables is as follows (Figure 6.6):

♢ **Decision level variables**:

$UG = \{0, 1, 2\}$ - user goal ($UG = 0$ - undefined goal, $UG = 1$ - the user is willing to see next exhibit, $UG = 2$ - the user is not willing to see next exhibit).

♢ **Modality events**:

$ORR = \{GB, no, yes\}$ - observed recognition result ($ORR = GB$ - GB answer, $ORR = yes$ - yes answer, $ORR = no$ - no answer);

$SMR = \{0, 1\}$ - speech modality reliability ($SMR = 0$ - unreliable speech recognition result: UG do not match ORR, $SMR = 1$ reliable speech recognition result: UG matches ORR);

$UR = \{0, 1\}$ - user in range ($UR = 0$ - user absent, $UR = 1$ - user present).

♢ **Modality features**:

$LSR \in \mathbf{R}^m$ - laser scanner reading (m is later defined to be 2);

$Lik \in \mathbf{R}$ - normalized frame likelihood;

$SNR \in \mathbf{R}$ - signal-to-noise ratio.

To find the optimal network topology and to evaluate the relative importance of the different aspects of the user goal we build the Bayesian network in two steps by modelling acoustic and spatial aspects of the user goal.

6.4.1 Bayesian networks for the acoustic aspects of the user goal

In the first step to infer the UG, information coming from the speech recognizer is combined only with information from the speech modality reliability aspect related to the reliability of the speech recognition.

Building the model

The subset $V_1 = (UG, SMR, ORR, Lik, SNR)$ of V is used in building the Bayesian network for speech recognition and speech modality reliability aspects (SMR, ORR) of the user goal UG. The observed variables in the experiment are the discrete variable ORR, and the continuous variables Lik and SNR. To account for the influence of the reliability aspect on the user goal we use the variable SMR.

In building the network topology we decide what will be the parent/children sequence following a top-down approach.

First, we order the set V_1 according to the level of significance for the UG classification task. We start with the decision level variable for the user goal UG as a root, then we continue with the modality events (user goal aspects): the SMR and ORR variables. At the end we finish with the observed variables ORR, Lik and SNR.

Second, we define the cause-effect relationships starting from the root variable and following the established ordering. UG is seen as the direct cause of the variables SMR and ORR and all the other variables, so we add the corresponding arcs. For example, if the goal of the user is to listen to the robot presentation, he will be staying in front of the robot pronouncing the word "yes", that in the ideal case would produce $ORR = yes$, $SMR = 1$ (reliable speech recognition result), and higher values for the Lik and SNR variables.

In the case of wrong recognition result we would expect that the small values for the Lik and SNR would provide evidence for unreliable recognition result $SMR = 0$ that can explain the uncorrect ORR. So we add tree arcs from SMR to ORR, Lik and SNR. We also assume that the ORR would influence directly the Lik and SNR. For example, when people are pronouncing, "yes" or "no" compared with the case when they remain silent would produce higher SNR and Lik values. That is why we add arcs from ORR pointing to Lik and SNR as well.

The topology of the Bayesian network BN1 fusing information from the acoustics-related aspects of the user goal (BN1) is given in Figure 6.6 (a). Shaded variables are observed during the inference of $P(X_Q|E) = P(UG|Lik, ORR, SNR)$, where $X_Q = UG$ and $E = (Lik, ORR, SNR)$.

6.4.2 Spatial aspect of the user goal

In the second step, the information coming from the speech recognizer is combined only with information from the spatial aspect related to the presence of a user.

The main decision variable is again UG. The observed variables in this case are the discrete variable ORR, and the continuous variables for the laser scanner reading and the likelihood: LSR

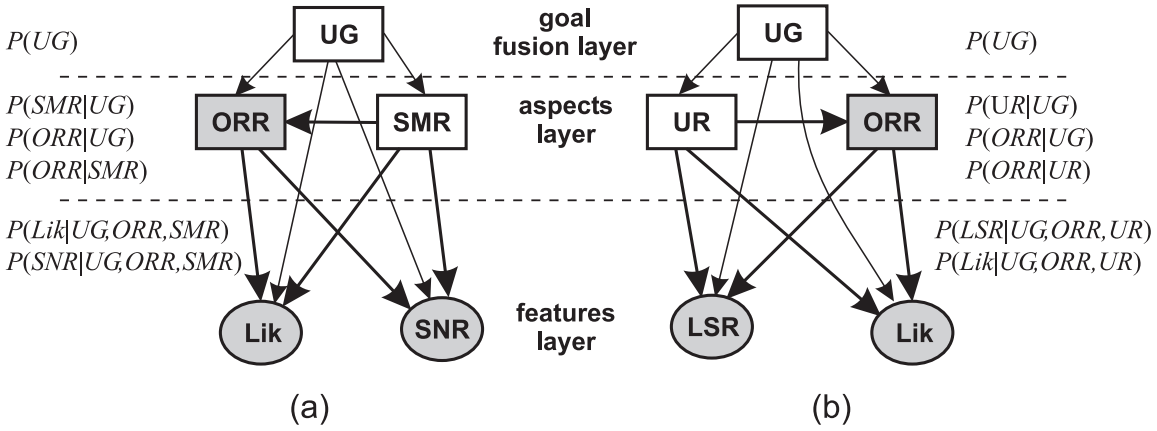


Figure 6.6: Bayesian network for (a) the acoustic aspects (BN1) and (b) the spatial aspect (BN2) of the user goal using ORR, SMR and UR variables

and Lik. To account for the influence of the spatial aspect of the user goal, we use the modality event UR representing the presence of the user in front of the robot.

Building the model

The ordered set of variables $V_2 = (UG, UR, ORR, LSR, Lik)$ is used in building the Bayesian network BN2 for the influence of the spatial aspect UR on ORR in inferring the state of UG.

In building the topology we use a top-down approach, where UG is seen again as the direct cause of the variables UR, ORR and the other variables (LSS, Lik), so we add the corresponding arcs. The presence of a user communicating with the robot given by the state of UR is seen as another cause for the particular values of the ORR as well as LSR and Lik.

There is an implicit assumption behind this statement that whenever there is a user near the microphone (UR=1) he is most probably speaking. Then his/her voice activity can affect the likelihood of the recognized words (Lik).

We assume that specific values of the $LSRandLik$ variables can be caused by particular words as given by the ORR, and we add two arcs from ORR to these two variables.

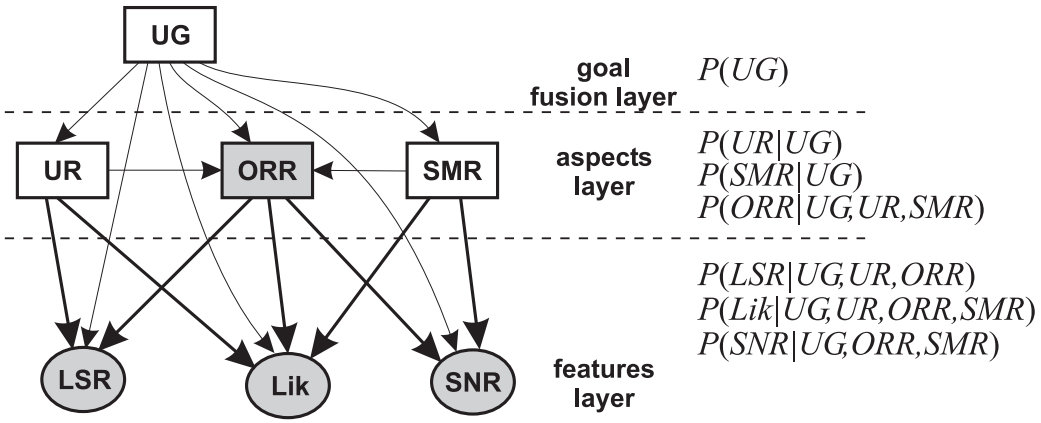
Figure 6.6 (b) depicts the topology of the Bayesian network BN2 built with the set of variables V_2 for the purpose of inference of $P(X_Q|E) = P(UG|LSR, Lik, ORR)$, where $X_Q = UG$ and $E = (LSR, Lik, ORR)$.

6.4.3 Combined topology

In the final stage we combine the two previous networks to account for all the user goal aspects in identifying the user goal UG. The Bayesian network is defined over the complete set $V = (UG, SMR, UR, ORR, LSR, Lik, SNR)$. The set of arcs comes from the two networks BN1 and BN2 that form the combined network BN. The particular structure already introduced in Figure 6.5 can be seen at the aspects layer in the final BN (Figure 6.7). It represents the inter-causal relations between UR, SMR and ORR as described in Section 6.3).

Figure 6.7 depicts the combined form of the Bayesian network built with the set of variables V for the purpose of the user goal classification. Shaded variables are observed during the inference of $P(X_Q|E) = P(UG|LSR, Lik, ORR, SNR)$, where $X_Q = UG$ and E= (LSR, Lik, ORR, SNR). According to the rules of d-separation (Chapter 4) all the observed variables (shaded ones) provide

evidence for UG in the current topology. In other words there are no "blocked" observed variables in the network.



Acronyms summary: UG - User Goal, UR - User in Range, ORR - Observed Recognition Result, SMR - Speech Modality Reliability, LSR - Laser Scanner Reading, Lik - Likelihood, SNR - Signal-to-Noise Ratio.

Figure 6.7: Combined Bayesian network for multimodal user goal identification

6.4.4 Training of the Bayesian networks

In order to perform consistent inference, the parameters of the Bayesian network CPDs (the conditional probability tables for the discrete variables and the parameters of the Gaussian pdfs for the continuous ones) have to be learned from data. In the case of full observability of the variables in the training set, the estimation can be done with random initialization and a maximum likelihood (ML) training technique. During the training the CPD parameters are adjusted in order to maximize the likelihood of the model with respect to the training data examples (Appendix C.2 in (Murphy, 2002)). The likelihood computation formulae needed to train the Bayesian networks used in our experiments are given in Chapter 4.

The networks from Figures 6.6 and 6.7 are used in the UG classification experiment. For training the models, we use 270 training examples for each value of UG, resulting in 810 sequences of the form: $\{UG, U, LS, UR, Lik, ORR, SNR\}$. We assume that the user goals have equal prior probabilities. The training data examples are taken from real data (audio files and laser scanner readings), collected during the deployment period of RoboX at Expo.02. The audio files contain a speech signal, sampled at 16 kHz, with duration of 2 seconds, corresponding to the average duration of yes/no answer. LSR vectors are calculated from the laser scanner readings generated by the scanner. The laser scanner reading (Figure 6.8) contains a sequence of values corresponding to the distances to the obstacles in the environment (walls, humans, etc.) reflecting the laser beam of the scanner. Within an angle interval of 360^0 and 0.5^0 resolution, the laser scanner reading results in 722 distances in meters (m) with resolution of 0.5 mm with respect to the robot (Jensen et al., 2002a). Only the values within the interval $[255, 285]^0$ are taken in order to account for presence of visitors in range for spoken interaction (the event UR). This angle sector corresponds to the front of the robot, where the microphone array is located. To eliminate noisy laser beam reflections and to reduce the dimensionality of the resulting vector, we divide this interval into two equal intervals, integrating the distance values contained in them, and normalizing the resulting values by the length of the intervals. The resulting two-dimensional vector $LSR = (d1, d2)$ is used as the variable LSR in the

Bayesian network.

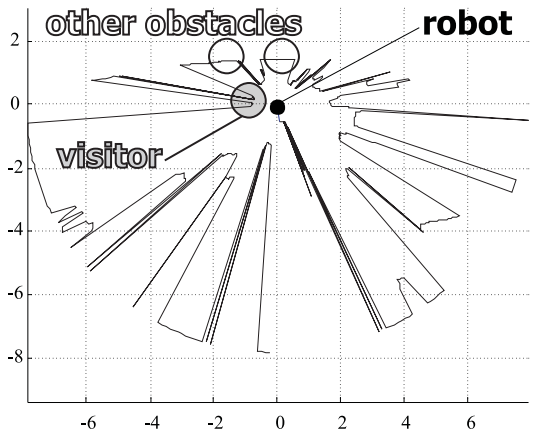


Figure 6.8: Laser scanner reading

ORR values are obtained after presenting the speech files to the recognizer of the robot. According to its definition, $SMR = 0$ when ORR does not match with UG and $SMR = 1$ when ORR matches the goal of the user UG. As already stated, $UR = 0$ corresponds to the event "there is no user in range for spoken communication" and $UR = 1$ corresponds to the opposite event. Hence, when $UG = \{1, 2\}$ then $UR = 1$, but when $UG = 0$ it may also happen that $UR = 1$. Finally, values for the SNR are estimated from the speech.

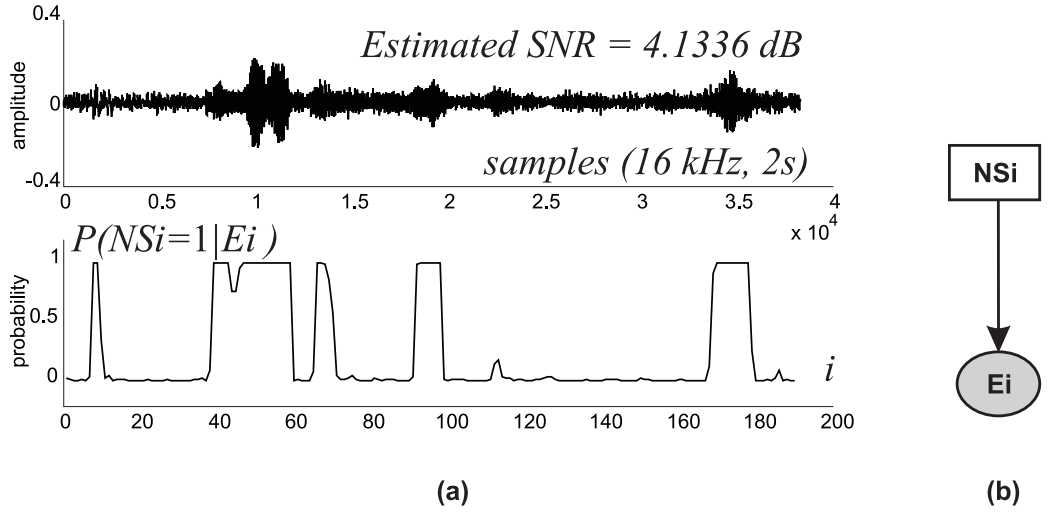


Figure 6.9: Experimental results (a) and BN (b) for SNR estimation

In order to estimate the real value for the SNR we need to separate the clean speech and the noise in noisy speech signal in the training data. This is not trivial, since in the noisy acoustic conditions of the exhibition, the signal from the visitor speaking to the robot can have similar characteristics to the background noise, mostly coming from other people speaking. Instead of performing costly calculations to separate speech and noise and calculate the real SNR, we estimate a SNR correlated feature based on the signal's short-term energy. Short-term energy is calculated using windows containing 400 samples (25 ms) with 50 % overlapping. We assume that each energy

value in this vector can be generated by two Gaussian distributions, modelling the probability of the current energy value being noise or clean speech segment in the signal. Such a model can be represented in the framework of Bayesian networks as shown in Figure 6.9 (b). NS_i is the hidden variable governing the current energy value being noise or speech, and E_i is the current energy value. This network is trained on the speech short-term energy vector, using the expectation maximization algorithm (EM) (Chapter 4) with random initialization. After training the model, we test it once again with the energy vector, inferring values for $P(NS_i|E_i)$, where $NS_i = 1$ corresponds to speech and $NS_i = 0$ to noise segments, for each energy component in the vector (Figure 6.9 (a)). The SNR correlated feature is defined as follows (Prodanov and Drygajlo, 2003):

$$SNR = 10 \cdot \log_{10} \left(\frac{\sum_i P(NS_i = 1|E_i) \cdot E_i}{\sum_i P(NS_i = 0|E_i) \cdot E_i} \right). \tag{6.1}$$

6.4.5 Testing of the Bayesian networks

For testing the models, we use 130 testing examples per given value of UG, resulting in 390 testing sequences that are independent of the training examples. Some statistics about the testing and training data including the averages and standard deviations (STD) per user goal and in total (for all the training and testing examples) for the two LSR components in meters, the recognition Lik and the SNR in dB are given in Table 6.1.

Training data statistics

UG \ Average	d1(m)	d2(m)	Lik	SNR(dB)
0 (GB)	4.02	4.66	-71.05	3.72
1 (yes)	2.52	2.47	-71.18	10.56
2 (no)	2.70	2.49	-70.81	8.98
Grand Total	**3.08**	**3.21**	**-71.02**	**7.76**
UG \ STD	**d1(m)**	**d2(m)**	**Lik**	**SNR(dB)**
0 (GB)	1.40	1.61	4.00	3.63
1 (yes)	1.79	1.69	3.41	3.66
2 (no)	1.72	1.65	3.73	4.37
Grand Total	**1.78**	**1.95**	**3.72**	**4.87**

Testing data statistics

UG \ Average	d1(m)	d2(m)	Lik	SNR(dB)
0 (GB)	3.75	4.43	-72.40	3.79
1 (yes)	2.41	1.70	-72.91	10.97
2 (no)	2.46	2.29	-72.46	9.33
Grand Total	**2.87**	**2.81**	**-72.59**	**8.03**
UG \ STD	**d1(m)**	**d2(m)**	**Lik**	**SNR(dB)**
0 (GB)	1.37	1.76	4.59	3.15
1 (yes)	1.83	1.31	3.54	4.57
2 (no)	1.52	1.45	3.78	4.65
Grand Total	**1.70**	**1.92**	**3.99**	**5.18**

Table 6.1: Data statistics

After training the networks, we perform inference on UG, given the evidence from the samples of testing data on LSR, Lik, SNR and ORR. Since our Bayesian networks have at most 7 variables, we use a method of exact inference based on the junction tree algorithm (Chapter 4). Using this algorithm, a value for $P(UG = ug|E = e) = P(UG = ug|ORR = o, Lik = l, SNR = sn, LS =$

$(d1, d2))$ is calculated for each $ug \in \{0, 1, 2\}$ and every testing sample $e = \{o, l, snr, (d_1, d_2)\}$. The result from the experiment for the combined Bayesian network (Figure 6.7) is depicted graphically in Figure 6.10. The first curve shows the true values for the UG for each testing sample. Values are sorted by the particular UG value for visual convenience. The other three curves show the values for $P(UG = ug|E)$, where $ug \in \{0, 1, 2\}$, inferred by the network. To select the most likely user goal we use a criterion similar to Equation 4.34:

$$\hat{ug} = \arg\max_{ug}(P(UG = ug|E = e)) \tag{6.2}$$

Results for the percentage of accurately classified cases, using the three Bayesian networks (BN1,BN2 and BN) are given in Table 6.2. The "ORR Acc" presents the accuracy of the speech recognizer on the audio part of the testing data. The rows "BN1 Acc, BN2 Acc and Final BN Acc" contain the accuracies derived from the three Bayesian networks (Figure 6.6 and 6.7) classifiers after calculating the corresponding $P(UG|E)$ and choosing a user goal (UG) according to the criterion (6.2). The accuracy is calculated by subtracting the number of UG misclassifications from the number of all testing samples dividing the resulting value by the number of the testing samples per user goal and for all user goals.

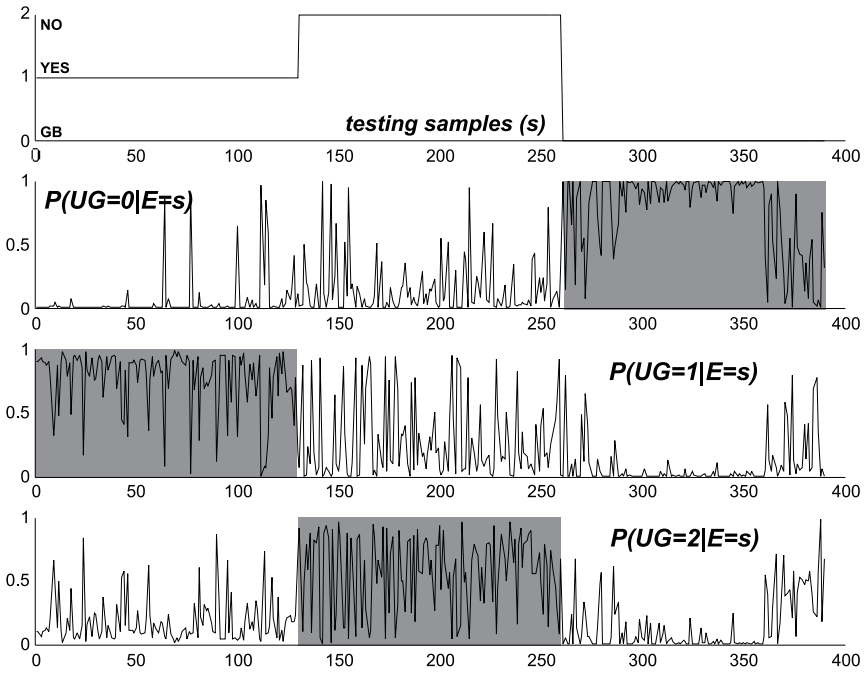


Figure 6.10: Graphical representation of $P(UG|LSR, Lik, SNR, ORR)$

6.5 Discussion

The results in Table 6.2 show a significant improvement in the accuracy of the user goal identification, when introducing information from the laser-related aspect and the speech recognition reliability aspect using a Bayesian network classifier ("Final BN Acc" in Table 6.2). The system can be used to avoid speech recognition errors without any dedicated repair dialogue technique. The gain in performance is due to the improved identification of the garbage case $UG = 0$, which in turn is due

UG	0 (GB)	1 (yes)	2 (no)	Overall
ORR Acc	**38.5%**	**93.1%**	**66.9%**	**66.2%**
BN1 Acc	77.7%	84.6%	61.5%	74.6%
BN2 Acc	78.4%	86.9%	66.9%	77.4%
Final BN Acc	**80.8%**	**85.3%**	**67.7%**	**77.9%**
Gain BN1	39.2%	-8.5%	-5.4%	8.4%
Gain BN2	39.9%	-6.2%	0.0%	11.2%
Gain Final BN	**42.3%**	**-7.8%**	**0.8%**	**11.8%**

Table 6.2: Experimental results for ORR and BN accuracy

to the dependencies found between the laser scanner data and the speech recognition result in the Bayesian network presented in Figure 6.7. According to the rules of d-separation, all the observed variables (shaded ones) provide evidence for revealing the state of UG in the topology in Figure 6.7. Evidence from the observed variables (ORR, LSS, Lik, SNR) can propagate following a direct path to the UG node as well as following paths through the unobserved variables UR and SMR. Thus the state of UG depends on both the observed values of the corresponding features as well as the inferred states of the unobserved variables UR and SMR. Hence, we have achieved both feature and decision-level (Figure 6.4) fusion in one pass using the Bayesian network. The observed testing results demonstrate the quantitative effect of the above presented dependencies. For example, in the region corresponding to the undefined user goal $UG = 0$ (the shaded region in the second top plot in Figure 6.10 the Bayesian network has calculated the following probabilities:

- $\diamondsuit$ $P(UG = 0|s_1) = 0.94, P(UR = 1|s_1) = 0.06, P(SMR = 0|s_1) = 0.06,$
  for the testing sample: $s_1 = (ORR = GB, LSR = (4.8, 4.6)m, Lik = -71.3, SNR = 7.8\text{dB})$;
- $\diamondsuit$ and $P(UG = 0|s_2) = 0.90, P(UR = 1|s_2) = 0.09, P(SMR = 0|s_2) = 0.94,$
  for the testing sample: $s_2 = (ORR = yes, LSR = (4.8, 4.1)m, Lik = -67.2, SNR = 1.2\text{dB})$.

It can be seen from the above testing samples that in both the cases people are far away from the tour-guide robot (more than four meters). In the first case the recognizer has correctly spotted garbage word GB, while in the second case there is an incorrectly recognized yes word. Despite the higher likelihood in the second case, the low probability of user presence - $P(UR = 1|s)$, and the low SNR value (giving rise to the probability of unreliable speech recognition result - $P(SMR = 0|s_2)$) provide evidence in favor of the right decision about the most likely user goal - $UG = 0$. The improved identification of the user goal can be used by the robot to acknowledge the absence of communicating visitor using the speech synthesis component. That kind of situation awareness would benefit to the quality of interaction, as well as the overall satisfaction of the visitors. As reported repeatedly in (Burgard et al., 1999; Drygajlo et al., 2003; Thrun et al., 1999a; Willeke et al., 2001), people find it very amusing when the robot is able to acknowledge awareness of their activities, such as blocking the free way of the robot, playing with its buttons etc. In our case the robot might ask for attention or simply stop talking when there are no visitors answering to it. The results presented in the third and forth row of Table 6.2 outline the relative importance of the additional information extracted from the UR (user presence) and SMR (recognition reliability) related data. It can be seen that introducing information from the laser scanner signal leads to greater benefits, compared with the case when only auxiliary information concerning the acoustic data reliability is used. In the case when $UG = \{1, 2\}$ there is not any gain in using the Bayesian network, which is an intuitive result as the laser scanner does not provide additional information for distinguishing between the spoken words yes and no. Additionally, when people are close to the robot and the models for speech recognition were trained with noisy speech conditions, the

results for yes and no can be unchanged or even slightly degraded. Possible accuracy improvement can be obtained using information from a video camera images tracking the lip-movement of the communicating speaker. Finally, the proposed error handling method can be easily applied in more complex dialogue systems employing keyword spotting based speech recognition systems. In particular using keywords associated with the particular user goals would not require any changes in the network topology. At the same time keyword recognition would avoid the additional complexity of the speech-understanding module. However, extending the model with additional modalities and user goals should be done after taking into account some important scalability issues, concerning the framework of Bayesian networks.

6.5.1 Scalability of Bayesian networks

First, the complexity of computing of exact inference in Bayesian networks with conditional Gaussians is NP hard (Cooper, 1990; Murphy, 2002). The "junction tree" algorithm used for inference in our case is done in two phases, i.e. constructing a junction tree from the original Bayesian network and performing inference on a junction tree after entering the evidence. The junction tree is a special undirected graph (Chapter 4), in which some of the original nodes in the Bayesian network are clustered together in order inference to be done in linear time with the number of nodes. The NP-hardness comes into place when the junction tree CPDs are constructed (Russell and Norvig, 2003). In our case we have a static Bayesian network, i.e. its topology remains unchanged during the different inference instances. In addition, the continuous variables in our case are observed, which avoids the problem of marginalizing continuous variables (Murphy, 2002). Thus the time of exact inference, once the junction tree is constructed is linearly dependent on the number of network nodes. The time complexity of constructing the junction tree with 3 user goals and 7 nodes, where the discrete variables are at most ternary in the worst possible case of fully connected graph is less than $O(3^7)$. Second, extending the network with additional nodes would require additional training data. Recording multimodal data in real time while the robot is interacting with people is a computationally demanding as well as a time consuming operation, since many interaction cycles will be required per given user goal in order to collect sufficient amount of training data. That is why deciding on an efficient, limited number of user goals is one important requirement for a real-time robotic application both from the usability and computational point of view.

6.5.2 Optimizing topology

In order to facilitate the task of inference we have done experiments to optimize the topology of the final Bayesian network (Figure 6.7). In these experiments some of the arcs without strong impact on the UG state were removed, e.g. the arcs pointing from UG to the continuous features (LSR, SNR and Lik). The comparison was done with respect to the overall accuracy of the combined Bayesian network ("Final BN Acc" in Table 6.2). The same training and testing data were used in the experiments as for the networks in Section 6.3.1. The overall accuracy of the user goal classification did not change after removing both the arcs pointing from UG to the SNR and Lik (Figure 6.11). After removing all the three arcs, including the one pointing at the LSR node the overall accuracy dropped by 1.02 %. This result supports empirically the intuitive hypothesis that fusion made at two levels at the same time (feature and decision level) might lead to greater benefits than just only at the decision level Smith (2003).

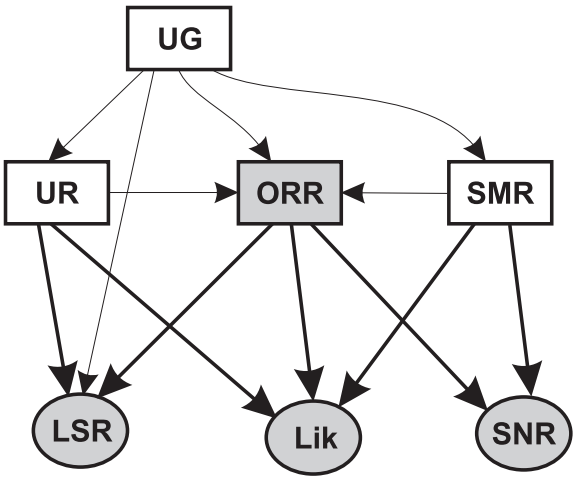


Figure 6.11: Optimized BN topology

6.5.3 Training data issues

Since the performance of speech recognition is decreasing in adverse acoustic conditions, it may not benefit substantially from additional speech training data. On the other hand, the use of additional data from acoustic insensitive modalities supplies auxiliary information not interpretable by the speech recognition system, but useful for detecting recognition errors in human-robot dialogues under adverse noisy conditions. The additional computational cost, required for exact inference with Bayesian networks, can be well compensated by the benefits from using such information for correcting speech recognition errors. If needed, faster algorithms for approximate inference (Jordan et al., 1999; Murphy, 2002) can be used with larger Bayesian networks, while incorporating additional user goals and modalities that may require more training data. In our experiments the Bayesian network training and testing data were taken from the modality data, collected during the deployment period of RoboX at the Expo.02. The size of the data set was chosen in order to clearly outline the benefit of using the additional laser scanner modality information, while keeping minimal amount of training data with an equal number of examples per given user goal.

6.6 Summary

In this chapter we introduced a new approach for error handling in spoken dialogue systems for mobile tour-guide robots working in mass exhibition conditions. The problem of dialogue management was shown to depend on a robust inference of the user goal at each dialogue state. While the process of identifying the user goal only from the speech recognition result can be inefficient in the noisy exhibition conditions, using the additional acoustic noise-insensitive laser scanner signal can be beneficial. The framework of Bayesian networks was introduced for detecting and correcting errors in the user goal classification problem using multimodal input. We demonstrated that a Bayesian network can model efficiently the dependencies between the speech and the laser scanner signals. In addition, the method allows for the explicit modelling of the speech recognition reliability enabling the possibility to exploit both the strengths and the weaknesses of the speech recognizer in deciding about the true user goal. The performance of the model was tested in experiments with real data from the database, collected during the deployment period of the tour-guide robot RoboX at Expo.02. The results show that the Bayesian networks provide a promising probabilistic framework

for error handling in multimodal dialogue systems of autonomous tour-guide robots.

While modality fusion can reduce the need for repair dialogues, repair actions are still needed in the case of undefined user goal in the robot dialogue. These undefined user goals often occur due to adverse acoustic conditions or uncooperative user behaviors. In such conditions, to avoid inefficient dialogues, the repair actions can also exploit non-speech based modalities (e.g. buttons input or "search for visitors" repair action).

Multimodal repair strategies in dialogues with service robots

7

In this chapter, we introduce dialogue repair methods that exploit the inherent multi-modality of the tour-guide robot, in order to reduce the risk of the human-robot communication failures. Bayesian networks fusing speech and other modalities during user goal identification serve as input to graphical models known as decision networks. Decision networks allow the definition of dialogue repair sequences as actions, and provide a utility-based decisions for selecting actions. The use of utilities allow the explicit modelling of preferences on repair actions that are efficient in the current interactive setting. The efficiency is related to fulfilling, in the limited time the task of the tour-guide robot to provide its user (visitor) with exhibit information. The benefits of the proposed repair strategies are demonstrated through experiments with the dialogue system of RoboX.

Defining a dialogue repair strategy, i.e. the succession in which input modalities and corresponding multimodal repair actions are processed, is fairly straightforward in the case of two modalities. However, introducing more modalities makes the design process cumbersome and calls for a systematic approach in order to enable modularity in the repair strategy design. Introducing new modalities in the user goal identification process can bring benefits in detecting and preventing possible communication failures during interaction (Chapter 6), however the used Bayesian networks can become complex and computationally expensive. Every new modality introduces new user-goal aspects and new modality events that have to be inferred by the Bayesian network. The new user goal aspects can enable specific repair actions, depending on the evidence that the modality event provides for a possible communication failure. Thus, increasing the number of input modalities raises questions related to the importance of the modality events and related user goal aspects (Chapter 6) for detecting communication failures, and the subsequent order in which the aspects' inference and repair action selections have to be performed. Ad-hoc repairs can result in inefficient time-consuming dialogue flow. Therefore, the provision of systematic approach in the repair strategy becomes important when the robot has to communicate with casual users in limited time.

In the second part of this chapter, we introduce a grounding state-based model to address the problem of systematic provision of dialogue repairs. The model is motivated by cognitive theories on how humans resolve communication problems in their dialogues. The model exploits the multiple modalities available in the robotic system to provide evidence for reaching grounding states. The proposed methodology is sufficiently generic to be applied in the general case of voice-enabled communication with service robots. The Bayesian network topologies, utilized in the grounding model, are specially designed for modularity and computationally efficient inference.

7.1 Repair strategies in tour-guide dialogue

In Chapter 6 we have used an *argmax* criterion on the posterior probability distribution inferred by the Bayesian network to decide for the user goal value (Equation 6.2). In our approach to dialogue modelling the user goal value at each dialogue state is used to select the next dialogue state, where each state has an associated dialogue sequence (sequence of scenario objects). In the case of undefined user goal ($UG = 0$), the dialogue sequence should be a repair sequence, i.e. a dialogue sequence dedicated to recover from unreliable speech recognition or the user behavior that could cause the undefined user goal. The *argmax* criterion ensures minimal error when predicting the user goal values after inference, however it may not be the best criterion when choosing the corresponding next state in tour-guide dialogue.

If the dialogue repair sequences are defined as actions that the robot can perform at each dialogue state, principles from decision theory provide explicit way of selecting actions, given the robot's actions preferences and the level of uncertainty in user goal identification at each dialogue state. Decision theory defines action selection strategies based on explicit measure of robot's action preferences named utilities and the principle of maximum expected utility (MEU) (Russell and Norvig, 2003). Different actions at different states in dialogue can have different utilities given the tour-guide task requirements (Section 7.3).

7.2 Repair actions and their utilities

In the context of utility driven tour-guide robot, the user goal values at each dialogue state can be preferred to a different extend by the robot. For example, the tour-guide robot might prefer the user goal $UG = 1$ (positive answer to a proposed service) than $UG = 2$ (rejection of the service). The tour-guide dialogue can be seen then as a process of decision-making, where at each state in dialogue a decision is made according to the evidence about the user goals and their associated preferences. The decision coincides with the "initiative/response" pair during which the robot probes the external environment and elicits a probability distribution over the robot's internal states, i.e. $P(UG|E)$.

The Bayesian network in Figure 6.7 from Chapter 6 can be used for inferring the user goal and the modality related events (UR - user in range for communication and SMR - speech modality reliability) combining multimodal information. With the help of this network we can compute the posterior distributions $P(UG|E)$, $P(SMR|E)$, $P(UR|E)$, where the set of observed variables is composed of the laser scanner reading LSR, the likelihood of the recognition result (Lik), the speech signal-to-noise ratio (SNR) and the observed recognition result (ORR), i.e. $E = \{LSR, Lik, SNR, ORR\}$. The posterior distributions can be associated with chance nodes in a corresponding decision network, making a Bayesian network an input for a MEU-based decision system (Chapter 4).

Then the actions with maximum expected utility in the case of the UG chance node, using the

MEU-principle, can be calculated as follows:

$$MEU(\hat{a}|e) = \arg\max_{a} \sum_{ug} P(UG = ug|E = e) \cdot U(ug, a). \tag{7.1}$$

If the robot behavior at each decision point in dialogue is governed by the MEU-principle, the robot is guaranteed to accumulate maximal utility at the end of the conversation.

In order to apply Equation 7.1, we still need to define precisely the set of robot's actions a and the concrete utility function (e.g $U(ug, a)$).

7.2.1 Defining actions and repair strategies

We define the selection of repair actions in the case of tour-guiding using the dialogue scenario presented in Section 6.2. The dialogue sequences presenting the exhibits in one complete tour are seen as valid dialogue actions for the case when the user is willing to see the offered exhibit ($UG = 1$). We will refer to these sequences as "present next exhibit" actions. On the other hand, the initiative/response pairs offering exhibit presentations to the visitors can be also seen as valid actions for the opposite case of $UG = 2$. We will refer to these actions as "offer another exhibit" actions.

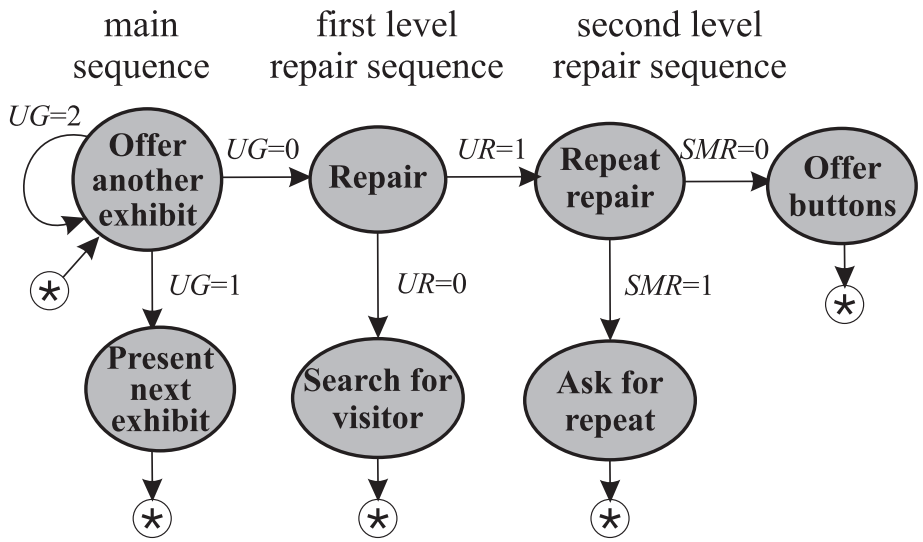
Due to uncooperative visitors and adverse acoustic conditions during dialogue, the visitor's intentions cannot always be classified into meaningful user goals in the context of tour guiding, e.g. simple accept/reject responses in the case of RoboX. In this case, using an "undefined" user goal ($UG = 0$) is well motivated and requires "repair" actions that the robot can perform to avoid communication failures. To define these "repair" actions, we take into account the tour-guiding dialogue requirements:

- ♢ Provide exhibit information through efficient speech-based interaction in limited time.
- ♢ The number of presented exhibits, after correct user goal identification, can be used as a measure for efficient interaction.

Defining repair actions and their succession

Dialogue repair sequences generally occur as an additional sequence in the normal process of human-robot interaction and may lead to delays in the communication process. Therefore, given the tour-guide dialogue task requirements the "repair" actions should avoid unnecessary repetitive patterns that might often arise using speech recognition in noisy acoustic conditions. In building "time-saving" repair sequences using alternative input and output robot modalities can be very beneficial. For example, in the case of absence of the communicating visitor ($UR = 0$, Figure 7.1) the most appropriate repair sequence should include an initial phase in which the robot moves around searching for a visitor. We will define such a repair sequence as the "Search for visitor" action. In the case of presence of a user ($UR = 1$), performing a "Repeat repair" action, e.g. asking the user for repeated input trial would be the fastest possible repair sequence. However, knowing that $UR = 1$ and $SMR = 0$ would give less motivation for the use of a speech-based "Ask for repeat" repair action, compared with an alternative use of the interactive buttons through the "Offer buttons" repair action.

The repair strategy outlined above is depicted in a form of state transition diagram in Figure 7.1. The state transition diagram for tour-guide dialogue represents a two-level repair strategy as outlined in the previous paragraph. In real conditions, however, the states of UG, UR and SMR are never known with certainty. If UG, UR and SMR are seen as chance nodes, decision networks can be



Acronyms summary: UG - User Goal, UR - User in Range for communication, SMR - Speech Modality Reliability.

Figure 7.1: Tour-guide dialogue state transition diagram

used as a state transition model for selecting valid actions using the principle of maximum expected utility (MEU), given by Equation 4.35.

7.3 Decision networks for tour-guide dialogue repair strategies

Figure 7.2 depicts the decision networks DN1, DN2 and DN3 that can be used for selecting actions in the three decision levels of the tour-guide dialogue in Figure 7.1. The Bayesian network from Figure 6.7 is used as an input for the three decision networks to output values for the corresponding posterior distributions needed for Equation 7.1, e.g. $P(S|E) = P(UG|E)$ in the main dialogue sequence case (DN1), $P(S|E) = P(UR|E)$ for the first level (DN2), and $P(S|E) = P(SMR|E)$ for the second level (DN3) of dialogue repair, given the evidence $E = \{LSR, Lik, SNR, ORR\}$ from the robot's input modalities.

The depicted decision networks utilize Equation 4.35 to compute the action with the maximum expected utility in each level in Figure 7.1.

In order to perform the computation we need to define the utility functions associated with the utility nodes in the three networks. These functions are defined as real valued tables, indexed by the actions and chance nodes.

In general, the numerical values of utilities are unique up to a positive affine transformation such that if $U(x)$ is the utility, then $k_1U(x) + k_2$ is equivalent for any constant $k_1 > 0$ and k_2 (Paek and Horvitz, 2003). The particular values in the utility tables corresponding to the three decision networks represent the tour-guide preferences about its actions, given the user goal values and are motivated by the tour-guide dialogue requirements presented in Section 7.2.1. These values can be interpreted as rewards that the tour-guide robot would gain in performing particular action, given the chance node values at the current decision point. For example, due to the time limit during interaction the most preferable action for a "rational" tour-guide robot would be to "present next

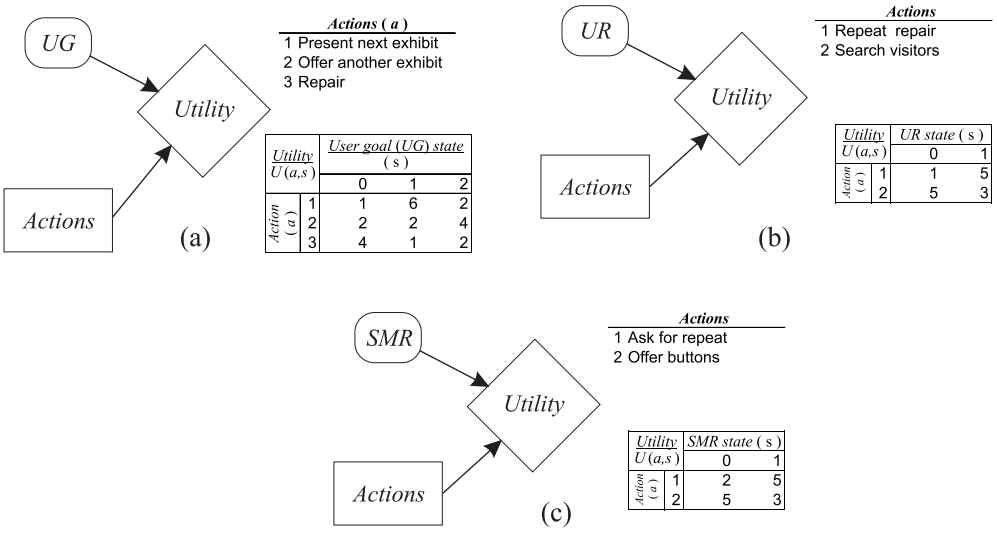


Figure 7.2: Decision network for managing the (a) main tour-guide dialogue sequence (DN1), (b) the first (DN2) and (c) the second repair (DN3) levels

exhibit" in the case of $UG = 1$, and the least preferable one would be the "Repair" action, since it might lead to not justified delays in interaction. However, in the case of $UG = 0$ performing the "Repair" action would be much more relevant in order to prevent communication failure. The above preferences are taken into account in the utility table in Figure 7.2 (a).

Given the utility tables, Equation 7.1 can be used by the three decision networks in the order specified in Figure 7.1 to select the actions that maximize the expected utility of that action, given the distribution over the values of the corresponding chance nodes (UG, UR and SMR).

7.3.1 Experiment with data from Expo.02

During Expo.02 we have collected multimodal data samples from the interactive tours of RoboX with the visitors (audio recordings and laser scanner readings, Chapter 6). The data were manually labelled with corresponding values for the user goal $UG = \{0, 1, 2\}$. Approximately 50 % of these samples were labelled with $UG = 0$. We have trained the BN in Figure 6.7 on a portion of 810 examples that resulted after balancing uniformly the UG values (270 examples for each user goal). Another balanced portion of 390 (130 examples per UG value) samples was used for testing the BN and results were reported in Chapter 6.

In order to outline the benefits of the proposed repair strategies, we have performed tests with only the data of $UG = 0$. We have used 130 testing examples (Figure 7.3) containing values of the three posteriors $P(UG|E)$, $P(UR|E)$ and $P(SMR|E)$ calculated by the BN in Figure 6.7 for 130 cases of an undefined user goal ($UG = 0$). The decision network DN1 was used initially to decide if a repair action is needed. In the case when the repair action had maximum expected utility, DN2 was used to decide if there is a visitor in front of the robot and consequently DN3 in order to decide what input modality has to be offered to the user during the repair sequence. The results from the experiment are shown in Table 7.1. The table depicts the correctness of MEU-based action selection at the main dialogue sequence and the proportion of selected repair actions in the first and second repair levels. Note that results are depicted only for the case of $UG = 0$ - no user in front of the robot. This case ideally requires the "Search visitors" repair action for all the 130 testing examples.

Comparative results for the overall UG identification task using $argmax$ criterion (Equation 6.2) and "MEU" based decision criterion are shown in Table 7.2.

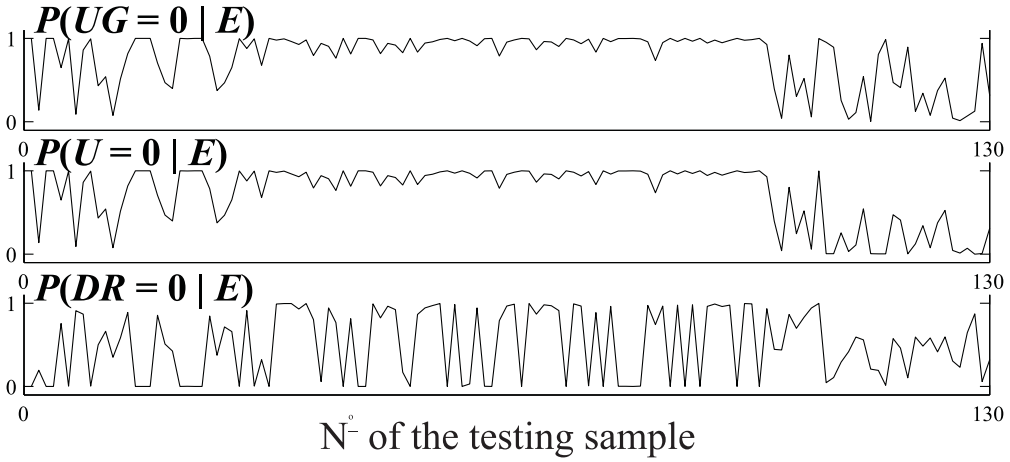


Figure 7.3: Graphical representation of the chance nodes' probabilities in DN1, DN2 and DN3 for 130 examples of $UG = 0$

Main sequence (DN1)		**MEU action**	**% corr.**
Actions:	1 Present next exhibit	9	6.9%
	2 Offer another exhibit	13	10.0%
	3 Repair	108	83.1%
1st level repair (DN2)		**MEU action**	**%**
Actions:	1 Repeat repair	6	5.6%
	2 Search visitors	102	94.4%
2nd level repair (DN3)		**MEU action**	**%**
Actions:	1 Ask for repeat	6	100%
	2 Offer buttons	0	0%

Table 7.1: Experimental results for $UG = 0$

As can be seen from Table 7.1 in 83 % of the cases the network DN1 has correctly assigned a repair action, and 94 % of the repair actions correspond to "Search visitors" actions. At the end, in all the 6 cases in which the user was estimated to be present, he/she is re-offered to use the speech modality during the final repair action. Finally, in 102 out of 130 cases the mobility of the tour-guide robot provides an efficient way to avoid communication failure due to the absence of visitor during interaction. We also see that among the wrongly selected actions at the main dialogue sequence level, the decision network DN1 has selected in most of the cases the "offer another exhibit" action. This decision can be seen as safer, compared with the "present next exhibit " action when there is no audience in front of the robot. This safer action selection strategy is explicitly encoded through the corresponding utility values in DN1 (Figure 7.2). Such repair strategies based on decision theory provide substantial degrees of freedom in modelling the tour-guide behavior. It can be seen from Table 7.2 that the performance of user goal identification does not change significantly when using $argmax$ or MEU criteria.

UG	Argmax Corr.	MEU Corr.	Argmax FAR	MEU FAR
0	80.8%	83.1%	6.5%	8.5%
1	85.4%	86.9%	15.4%	15.8%
2	67.7%	63.8%	11.2%	8.8%
Mean:	**77.9%**	**77.9%**	**11.0%**	**11.0%**

Legend: **Corr**. - Correctness is computed as the percentage of the correct identification out of all examples for a given UG value in the test data; **FAR**. - False Alarms Rate for a UG value is computed as the percentage of the identifications falsely assigned to the given UG value within all test examples labelled with UG values different from the given one.

Table 7.2: Correctness (Corr.) and false alarms rate (FAR) of UG identification using $argmax$ and MEU criteria on $P(UG|E)$

7.4 On the role of utilities and different modalities in the repair strategy

7.4.1 Global preferences on actions

Given equally likely chance node values (maximum uncertainty in the chance node distribution) the MEU principle will select the action with the maximal sum of the utilities across all user goals (the sum of the rows in the utility tables). In that sense the individual $U(a, s)$ values also contribute to the global preference on actions. Following such global preference the behavior of the tour-guide robot during interaction can be adapted to be more conservative or less conservative in performing the repair actions. For example, in Figure 7.2 (a) the global preference for presenting exhibits is higher compared to the one for offering a new exhibit or the repair option. Since searching for visitors might encourage the visitors around the robot to join the interaction, the global preference is in the favor of the "Search visitors" action in the first level of the tour-guide repair strategy (Figure 7.2 (b)). In the decision network corresponding to the second repair level (Figure 7.2 (c)), i.e. "Ask for repeat" vs "Offer buttons", the second action can be seen as globally more preferable. Since button's input during speech-based interaction does not depend on the acoustic noise, it is considered as more reliable at high levels of acoustic noise.

7.4.2 Executing repair actions over time

Given that visitors might utter out-of-vocabulary words, the "Ask for repeat" action may lead to delays in conversation. To handle this issue making the utilities dependent on the number of times an action is executed (e.g. $U_t < U_{t-1}$) might be beneficial. In other words, whenever we encounter a repeating repair action in a repair session, we can reduce its utility with respect to the utilities of the alternative actions. In this way we give a better chance to these alternative actions that can be more efficient in the current decision point, given the time requirements of tour-guiding. For example, in the second level repair, the buttons action can require less time than the alternative "repeat repair" action. The interactive buttons explicitly limit the decision choice of the user, who might be willing to play with out of vocabulary words.

It is also a good practice to equip the repair actions with an execution timeout. The timeout is needed by unpredictable situation in which the repair action will fail to produce an outcome that will normally result from normal user behavior (e.g. user found "UR=1" after "search visitor" repair, or "button pressed" after "offer buttons" repair). For example, if the "search for visitor" repair action is executed without any visitor in the exhibition room, it will be inappropriate for an "intelligent"

tour-guide robot to continue infinitely with the repair activity.

In all cases, a timeout on repair execution signals a repair failure. The repair failure should lead to reducing the preference on this action in a possible future repair. Such behavior of the robot can be interpreted as the "act of loosing interest in repetitive actions" that can be modelled through manipulating the utility values. In that way, in the next decision point the tour-guide will be more interested in actions that have not been recently tried for resolving the problem (e.g. "offer buttons" instead of "ask for repeat"). In modelling preferences in time by reducing the utility value, different functions can be used. For example, in economics the utility of the amount of money that a gambler would bet has been found to change according to the logarithm of the total amount of money the gambler possesses (Russell and Norvig, 2003). If a repair failure can be seen as a lost bet, the reduction in the interest of executing failed repairs can be modelled by the reduction in the logarithmic utility function of money, corresponding to fixed amounts of money.

In order to save time, a repair action has to be also executed a fixed number of times. The above mechanism can be modelled by the utility framework. Let us assume that all the repair actions in Figure 7.1 are resulting in failures (e.g. $UG = 0$ followed by $UR = 0$ and timeout or $UG = 0$, followed by $UR = 1$, $SMR = 0$ and timeout). In this case the utility values in the utility tables for DN2 and DN3 will begin to gradually decrease. Given a proper initialization, the utility values can fall below 0 and this can be an indication of a failure of the whole repair session. In this case, the tour-guide robot can suspend his current activities, where a pressed button or detected user in range for communication ($UR = 1$) can serve as a wake up signal. On waking up, the utility tables can be set to their initial values.

When initializing the utility table, two special cases can be of interest. First, the use of identity matrix for the utility table in a MEU-based decision system (Equation 7.1) is equivalent to using an *argmax* criterion on the chance node posterior distribution (e.g. Equation 6.2). Second, if all entries in the utility table are equal to 1 or to a constant number, the MEU-based system will result in equal expected utilities for all actions defined in the decision network. Thus, all decisions are equally attractive for the tour-guide robot or in other words it will have equal preferences for all actions.

Using the above special cases, one possibility for initializing the utility tables in DN2 and DN3 (Figure 7.2 (b) and (c)) can be the identity matrix. The state model in Figure 7.1 can be used then for executing repair actions and utility values can be manipulated over time as it was described above. The repair suspending criterion can be triggered when the utility table of one of the decision networks in Figure 7.2 is having all values less or equal to zero.

7.4.3 Incorporating new modalities and repair actions

The repair strategy presented in Section 7.3 relies on two modalities and relatively small amount of repair actions (Section 7.2). However, we may need more than the laser scanner evidence to assess the state of user attending to the conversation with the robot. For example, obstacles similar in shape and form to the legs of people can produce false user detection. In such cases a face detected in the video modality combined with evidence from the laser modality of the robot can result in a more robust user detection.

Incorporation of new modalities in the Bayesian network for user goal identification would result in additional nodes in the model. In the fusion method, using BNs presented in Chapter 6, these nodes correspond to modality-related events that reveal new user goal aspects of the final user goal. New nodes have to be introduced for the new modality features as well. The feature variables provide evidence for the modality event, where the rules of evidence propagation in the network are defined by the arcs that account for dependencies among the modality features, the modality-related

events and the final user goal. Without any topological restriction the Bayesian network can become computationally expensive, as probabilistic inference becomes NP-hard in multi-connected Bayesian network (Jordan et al., 1999). Therefore, topology restrictions that can lead to efficient inference are worthwhile investigating when constructing Bayesian networks for multimodal repair strategies for speech-based interaction with robots.

New modality events are typically associated with repair actions. These actions are to be executed when the event node probability provides sufficient evidence in favor of a specific failure (e.g. missing user in range for communication, missing face, etc.). When the robot operates with fewer modalities and repair actions, the repair strategy is straightforward to implement as in the case presented in Figure 7.1. However, incorporation of new actions and modalities will increase the possibilities for the repair action sequence. Therefore, a systematic approach for modality event monitoring and failure prediction will be needed in order to design a repair strategy consistent with the requirements of human-robot interaction.

Systematic approaches follow established methods in contrast to ad-hoc procedure in building the dialogue repair strategy during human-robot interaction. The repair strategy defines the sequential order in which triggering modality events has to be monitored and corresponding repair actions has to be executed. The final goal of the repair execution schedule will be to reduce the risk of communication failures in the process of spoken interaction. In the above context, strategies for dialogue repair that people typically use in their conversations are appealing to the human users, and are worth investigating in the repair strategy design.

7.5 Grounding in service robot human-robot spoken interaction

When designing conversational systems for service robots we have to be aware that misunderstandings about the communication goals of the participants occur even in conversations between humans that are thought to have ”perfect” speech recognition abilities. If not handled, these misunderstandings might result in communication failures. In the case of a conversation between people, misunderstandings are collaboratively resolved by the dialogue participants. People coordinate their individual knowledge states by systematically seeking and providing evidence about what they say and understand, which is known as the process of grounding in conversation (Clark and Schaefer, 1989). The amount of effort that people spend to ground their conversation at each dialogue turn is governed by a grounding criterion. The grounding criterion is used to evaluate the level of understanding between the dialogue participants. It is used to evaluate if the level of understanding in dialogue is sufficient for the current dialogue purpose, or if there is a risk of misunderstanding. In a service robot dialogue the grounding criterion can be related to the strength of evidence needed for identifying a particular communication user goal. The strength of evidence about the user goal can be quantitatively estimated by the posterior probability of the user goal given the evidence contained in the modalities' data. One of the sources for such evidence is the participant's feedback, another source can be the environmental conditions. For example, in very noisy acoustic conditions a speaker will specially seek the attention of the listener by looking him in the eyes, using much louder voice and repeating the important terms waiting for an appropriate acknowledgement. On the contrary, in quiet conditions all of these actions might slow down the interaction and even frustrate the listener. Hence, detecting a stronger evidence of adverse acoustic conditions should normally be one of the parameters used by the grounding criterion threshold, given that dialogue participants want to understand each other.

The dialogue participants in a service robot dialogue are the robot and its user. The user is

the person staying usually closest to the robot's front, communicating with the robot using speech. Most of the service robot applications take place in open spaces, where speaking people other than the user and the robot equipment itself can contribute to high levels of noise in the acoustic space. The speech in the input audio signal can originate from the user, but also from other people speaking (passers by) causing errors in speech recognition. Additionally, the end users of service robots can be ordinary people lacking any prior experience with robots. In the case of tour-guide or shop-assistant robots, users can decide to leave the robot at any time, since this type of interaction is typically short-term. Moreover, earlier work has pointed out cases when the users even try to confuse the robot for fun, e.g. misbehaving visitors in a tour-guiding scenario (Drygajlo et al., 2003; Willeke et al., 2001). Such behaviors make users' intentions difficult to anticipate in human-robot interaction, causing ambiguity and errors when the robot has to interpret them. Communication failures may arise in dialogue due to the above outlined factors. Hence, a service robot managing spoken dialogue with people needs to establish sufficient level of grounding with its user for minimizing the risk for communication failures. A sufficient level of grounding would mean that the robot has obtained sufficient evidence that the following grounding states have been reached: (1) user is attending to the conversation and (2) the speech modality is reliable in the current acoustic conditions.

In human-robot interaction, evidence for reaching grounding states can be delivered by information from speech as well as other modalities available on the robotic platform. For example, the state that the user is attending to the conversation can be revealed through her/his voice activity, combined with information from the video modality. If the robot asks the user for a repeated trial in which even alternative input such as buttons can be used, the unreliable speech recognition in very noisy conditions can be avoided. To ensure such functionality the robot needs a model to infer the corresponding grounding states such as the state of attending user or the state of speech modality reliability related to unreliable recognition. Since the end-users behavior can vary largely during their communication with the robot and the acoustic conditions are *a priori* unpredictable, the corresponding grounding states can be never inferred with certainty. Moreover, the limitations of the current sensor technology that is prone to measurement errors can lead to imprecise modality information. Hence, models based on deterministic mapping between input modality features and corresponding grounding states and user goals can lack sufficient robustness to the uncertainties of real-life service robot dialogue. Probabilistic models can deal with uncertainty using parametric models of distributions over random variables. The random variables can be associated with the grounding states and features derived from the robot modalities. The relations between the grounding states and their corresponding modality features can be seen as causal relations. Bayesian networks are widely accepted framework for efficient modelling of the probability distribution over a set of random variables by encoding the independence assumption behind the variables' causal relations. Hence, we use Bayesian networks for grounding modelling of spoken interaction between a user and a mobile service robot in mass exhibition conditions (tour-guide robot). While incorporating information from additional modalities can bring benefits (Prodanov and Drygajlo, 2005) in detecting possible communication failures during interaction, the resulting model that should infer grounding states and user goals using Bayesian networks can become complex and computationally expensive. Hence, providing Bayesian network topologies that allow straightforward incorporation of new modalities in the grounding model and computationally efficient inference becomes important.

7.6 Multimodal grounding in service robot dialogue

To build the grounding model for speech-based interaction between a user and a service robot, we take inspiration from the state model presented in Table 3.1 (Chapter 3).

7.6.1 Grounding states in human-robot interaction

We adapt the original model with the grounding states needed by a "collaborative" service robot in order to decide if the input audio signal is sufficiently grounded, relying on information from speech and non-speech modalities. The modified multimodal grounding state model is depicted in Table 7.3.

To avoid interpreting background noise as user input, the service robot has to be able to distinguish the potential user from people that are not using the system. It should have positive feedback from the user for reaching grounding states S0 and S1 in Table 7.3. Interested and collaborative users provide positive feedback showing attention by looking at the robot. To facilitate collaborative communication, the devices of the service robot are typically arranged to mimic anthropomorphic elements (e.g. a mechanical face), where a camera is typically located (Figure 5.1) Jensen et al. (2005). A collaborative user is assumed to stay close to the robot (S0 reached) looking at the robot's "face" (S1 reached) while communicating the user goal. A correct user goal interpretation using speech recognition requires that the speech recognition result is reliable (S2 reached), where the speech recognition reliability is mostly affected by the level of the background acoustic noise (Huang et al., 2001). To be understood by the robot, the user request has to be interpreted as a valid user goal, i.e. a goal that can be mapped into an existing service offered by the robot (S3 reached). Similarly to the original model (Table 3.1), reaching all the states in Table 7.3 signifies that the user speech input is grounded (understood by the robot) for the purpose of the service robot task oriented dialogue.

State	Modality / Event	Description
S 0:	Laser / $UR = 1$	***U***ser present in ***R***ange for communication
S 1:	Video / $UA = 1$	***U***ser ***A***ttending (looking at the robot)
S 2:	Speech / $SMR = 1$	***S***peech ***M***odality is ***R***eliable
S 3:	Speech / $UG \neq 0$	Robot identified a valid ***U***ser ***G***oal

Acronyms summary: UR - *U*ser in *R*ange, UA - *U*ser *A*ttending, SMR - *S*peech *M*odality *R*eliability, UG - *U*ser *G*oal.

Table 7.3: Multimodal state model of grounding in human-robot conversation

Failure or success to reach a given state is signaled by the evidence provided in the information from the robot's input modalities, such as speech, video, laser, etc. Information is extracted out of each modality in the form of events that can be inferred from the raw modality data. For example, the binary event "$UR = 1$" that a user is staying in close range in front of the robot can be inferred from the information contained in the laser scanner data. The binary event "$UA = 1$" - "User attending" can be inferred from information extracted from the video modality for a presence of a frontal face in the camera view. The event "$SMR = 1$" corresponding to "speech modality is reliable" can be inferred from information from the speech modality and the level of acoustic noise. $SMR = 0$ means that there is an error at the output of the speech recognizer (see Section 7.7.2 for more details). Finally, the speech modality is used to identify the user goal defined by the event UG, where $UG = 0$ means an undefined user goal and $UG \neq 0$ means a "valid" user goal, i.e. a goal that can be mapped onto existing robot-provided service. Examples of valid user goals are presented in Section 8.3.1. The events and their association with the grounding model states are depicted in Table 7.3.

Whether a grounding state is reached, directly depends on the strength of evidence for the events

as provided by the information from the input modality data. Given that the last grounding state is reached ($UG \neq 0$) would mean that S2 has been reached too ($SMR = 1$), which in turn means that S1 is reached ($UA = 1$) and S0 is reached ($UR = 1$), since an attending user implies a user who is close to the robot. All the above states and the propagation of evidence about their possible instantiations can be modelled by a Bayesian network. Then the strength of evidence about the modality related events can be quantitatively estimated by the posterior probability of the event given the evidence from the modality data, for example, the posterior probability of the event "valid user goal": $P(UG \neq 0|E = e)$, for the variable UG in the Bayesian network given the evidence $E = e$ from the input modalities. The posterior probabilities over the grounding states can be used in the grounding criterion in the case of service robots. The criterion can be formulated in the following way: in order to consider a grounding state as reached, the posterior probability of the corresponding modality event (e.g. $P(UR = 1|E)$) should be above chance level (above 0.5 in the case of a binary modality event). Thus the posterior probability below chance level (e.g. $P(UR = 1|E) < 0.5$) signify possible failures to reach a particular state in the grounding model that will require corresponding grounding (repair) actions.

In building the grounding model for service robot dialogue we use the mobile tour-guide service robot RoboX (Figure 5.1) as an example.

7.6.2 Two-phase grounding for user goal identification

The speech modality of RoboX is the main modality used for inferring the goal of the user out of the possible goals defined at each particular dialogue turn (Chapter 6). The User Goal (UG) is derived from the spoken user request for a service during the speech acquisition phase. In order to minimize the possible communication failures, user goal inference is performed in two consecutive phases in the multimodal grounding model.

- ♢ In the first phase (S0 and S1), the robot requires sufficient level of grounding as far as the user attendance to the conversation is concerned. Sufficient level of grounding requires strong evidence that the state S1 is reached, which also implies that S0 is reached (Table 7.3). This is needed for the robot to proceed to the second phase.

- ♢ In the second phase (S2 and S3), the robot seeks for sufficient level of grounding as far as the speech modality reliability is concerned. This would mean that state S2 is reached, after which S3 can be evaluated from the speech recognition result.

The reason behind the phase definition stems from the fact that it does not make sense to check the modality reliability and infer a user goal, if the user is not there, or is not paying the needed attention in the conversation. In that cases the user goal UG can be set to the undefined goal (UG=0). Only after achieving the two phases of grounding, the robot can reliably identify user goals from the underlying speech modality. The two phases for inferring user goals are depicted in Figure 7.4.

The grounding states and their associated modality events are depicted in the figure along with arcs indicating the causal relations between them as well as the corresponding modality features. LSR denotes the laser scanner reading, which is supplied by the laser modality. FD denotes the face detection ($FD = 1$ a face has been detected in the current video data, $FD = 0$ no face in the current video data) that is a binary feature derived from the video modality . ORR corresponds to the observed recognition result (recognized keywords) supplied by the speech modality. In Figure 7.4, the modality-specific events (e.g. UR (user in range) - laser, UA (user attending) - video) can be seen as the causes behind the particular input observations (feature values - LSR (laser scanner

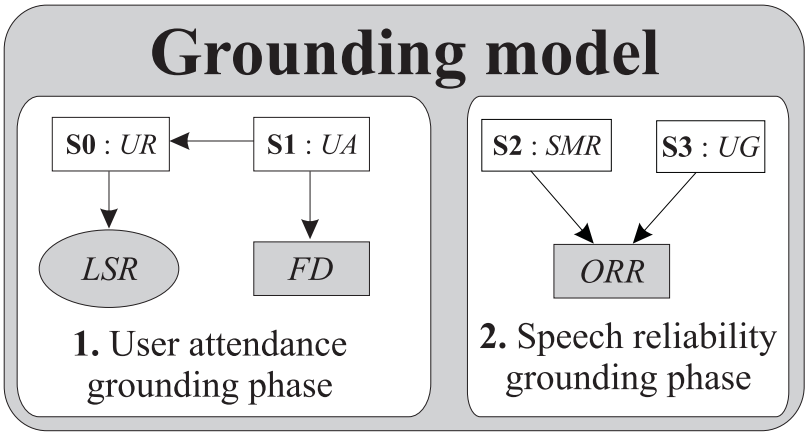


Figure 7.4: Two-phase grounding architecture for reliable speech-based UG identification.

reading) - laser, FD (face detection) - video). Through its events every distinct modality provides information about a particular aspect of the user goal (UR - laser, UA - video). The final user goal can be causally related to specific instances for all modality specific events. For example a valid user goal ($UG \neq 0$) would be causing $UR = 1$ and $UA = 1$. Inferring the user goal in multimodal system can be possible only when fusing information from one or more of the input modalities. Thus, fusing the different user goal aspects, as represented by the possible instantiations of the modalities' events can result in more robust user goal identification, compared with using only one modality (Prodanov and Drygajlo, 2005). In the multimodal fusion, we have to take into account the fact that the modality events are not deterministically related with the underlying modality features. For example, the recognition result (ORR) is affected by the ambient acoustic noise as well as the intra- and inter-speaker variability of speech. Hence, the cause-effect relation between the user goal and the speech recognition result should be seen as probabilistic. This argument is valid for the other modalities as well, i.e. laser and video.

7.7 Bayesian networks for grounding

In this section we use Bayesian networks for building the two-phase grounding model for user goal identification in service robot dialogue (Figure 7.4).

7.7.1 Bayesian network for the attendance grounding phase

The Bayesian network for the first phase of grounding is depicted in Figure 7.5 (a). It contains two discrete variables UR and UA corresponding to the events "User in range" for communication and "User attending" associated with the grounding states S0 and S1. These variable have direct causal impact on corresponding features derived from the laser and video modality that are represented by the two observed variables LSR and FD. LSR is a continuous variable corresponding to the laser scanner reading. Each raw scanner reading contains samples within range of 360^0 with precision of 1^0. The samples correspond to the distances from obstacles that reflects the laser beam or to the nominal range of the laser range finder which is 9 m. In order to extract features for detecting legs in the sequence of distance samples certain preprocessing steps are needed. Details concerning the preprocessing step performed on LSR for leg-detection can be found in Section 8.3.3. FD is a binary variable corresponding to a video modality feature indicating a face detected in the video

stream (FD=1). Finally the event of "User attending" (UA=1) to the conversation is seen as the cause of the event "User present" ($UR = 1$). In the first phase of grounding, the full set of variables is $V = (UA, UR, LSR, FD)$. Taking into account the arcs defined in Figure 7.5(a), the joint pdf over V can be written as:

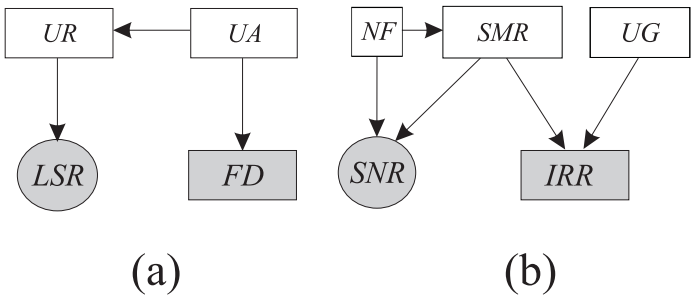
$$P(V) = P(UA)P(FD|UA)P(UR|UA)P(LSR|UR). \tag{7.2}$$

The first grounding phase is reached in the case, when $UA = 1$. The criterion for engaging in a grounding action at this phase is based on the posterior probability $P(UA = 1|E)$, where the set of observed (evidential) variables contain LSR and FD in this case, i.e. $E = \{LSR, FD\}$. Given the BN topology, the posterior distribution over the binary variable UA is calculated by the formula:

$$\begin{aligned}\mathbf{P}(UA|lsr, fd) &= \alpha \sum_{UR} P(UA)P(fd|UA)P(UR|UA)P(lsr|UR) \\ &= \alpha P(UA)P(fd|UA) \sum_{UR} P(UR|UA)P(lsr|UR),\end{aligned} \tag{7.3}$$

where $\mathbf{P}(UA|lsr, fd)$ denotes a two component vector, and $e = \{lsr, fd\}$ corresponds to the particular instantiations for the evidence variables LSR and FD. Particular UA value is chosen applying the $argmax$ criterion on the posterior probabilities defined by Equation 7.3:

$$\hat{ua} = \arg\max_{ua}(P(UA = ua|E = \{lsr, fd\})). \tag{7.4}$$



Acronyms summary: UR - *U*ser in *R*ange, LSR -*L*aser *S*canner *R*eading, UA - *U*ser *A*ttending, FD - *F*ace *D*etected, UG - *U*ser *G*oal, SMR - *S*peech *M*odality *R*eliability, NF - *N*oise *F*actor, SNR - *S*ignal-to-*N*oise *R*atio, IRR - *I*nterpreted *R*ecognition *R*esult.

Figure 7.5: Attendance grounding phase (a) and speech reliability grounding phase (b) BNs

7.7.2 Bayesian network for the speech reliability grounding phase

In the second phase of grounding, the user goal is inferred after ensuring that speech modality is reliable. The level of speech modality reliability is related to the probability of the event of mismatch between the true user goal value UG and the one obtained from the observed recognition result (ORR). We denote the user goal value obtained from the ORR as IRR (Interpreted Recognition Result). Given the definitions provided in Chapter 6, we can write that if ORR = GB then $IRR = 0$ if $ORR = keyword1$ then $IRR = 1$, if $ORR = keyword2$ then $IRR = 2$, etc. For example, in the case of $ORR = \{GB, yes, no\}$, $IRR = \{0, 1, 2\}$. Then, the event of mismatch between UG and IRR can be written as ($UG \neq IRR$). To define the reliability measure we introduce a binary variable SMR, where $SMR = 1$ represents the event "speech modality is reliable" ($UG = IRR$) and

$SMR = 0$ represents the opposite event, i.e. $(UG \neq IRR)$. The Bayesian network in Figure 7.5 (b) depicts a causal model for the variables UG, IRR and SMR. In this network the user goal value can be seen as the cause of the particular interpreted recognition result, and the speech modality reliability can be seen as an alternative cause that might also point at errors in the IRR value. For example, $IRR = 1$ can be caused by $UG = 1$ and $SMR = 1$ (the speech modality is reliable) or $UG \neq 1$ and $SMR = 0$ (the speech modality is unreliable). Since the variables UG and SMR are not observable during the conversation with the robot, we need to provide additional sources of information that can be observed and can provide evidence in favor of the particular UG and SMR values. The "noise factor" NF, which corresponds to the event of high level of acoustic noise can have strong causal impact on the SMR variable. A signal quality measure can be used to provide evidence for the NF variable. For example the signal-to-noise-ratio (SNR) of the speech signal can be used to account for the level of acoustic noise in the speech modality, which is known to be one of the main degradation factors for the performance of the speech recognition systems. Therefore, we define the variable $NF = \{1, 0\}$ corresponding to the binary event of "high/low level of acoustic noise" which has causal impact on the continuous variable SNR. SMR, can be also seen as a cause for particular SNR values. Given the BN variables set $V = (UG, SMR, NF, IRR, SNR)$, and taking into account the arcs defined in Figure 7.5(b), the joint pdf over V can be written as:

$$P(V) = P(UG)P(NF)P(IRR|UG, SMR)P(SMR|NF)P(SNR|SMR, NF). \tag{7.5}$$

The posterior $P(SMR|IRR, SNR)$ is the distribution of the speech modality reliability measure. Following the network topology the posterior probability over SMR can be written as:

$$P(SMR|irr, snr) = \alpha \sum_{UG,NF} \Big(P(UG)P(NF)P(irr|UG, SMR) \cdot \; P(SMR|NF)P(snr|NF, SMR) \Big)$$

$$= \alpha \sum_{UG} \Big(P(UG)P(irr|UG, SMR) \Big(\sum_{NF} P(NF)P(SMR|NF)P(snr|NF, SMR) \Big) \Big), \tag{7.6}$$

where $\{irr, snr\}$ correspond to the particular instantiations for the evidential variables in the Bayesian network. In the second row we apply the distributive law in order to avoid unnecessary computations (Aji and McEliece, 2000). We have defined the event $SMR = 1$ as the indicator of the event $(UG = IRR)$. Then, given that $SMR = 1$ the probability values for $P(IRR = irr|SMR = 1, UG)$ become $P(IRR = irr|SMR = 1, UG = irr) = 1$ and 0 for the rest UG values. In this case, the Equation 7.6 can be simplified in the following way:

$$P(SMR = 1|irr, snr) \propto P(UG = irr) \sum_{NF} P(NF)P(SMR = 1|NF)P(snr|NF, SMR = 1), \tag{7.7}$$

where $\propto$ is the proportionality symbol. Since all the entries for the probabilities $P(IRR|SMR = 1, UG)$ are zero except the case of $UG = IRR$, the summation over UG in Equation 7.6 is reduced to the multiplicative term $P(UG = irr)$. This leads to a reduction in the number of operations needed for computing $P(SMR = 1|irr, snr)$. The above formula shows that the probability of reliable speech modality given values for the observed recognition result and the SNR is proportional to the prior probability of the user goal value corresponding to the particular observed IRR value multiplied by a weighted sum of two Gaussian components. These components correspond to the probability of the observed SNR given the noise factor value and $SMR = 1$. The probability is weighted by two weight components, i. e. $w_1 = P(NF)$ - the prior probability of each NF value (the prior probability of high level of noise), and $w_2 = P(SMR = 1|NF)$ - the causal impact of the

noise factor on the event ($UG = IRR$). The likelihood $P(snr|NF, SMR = 1)$ can be also seen as a measure of the strength of evidence of noise after observing the acoustic environment (the current SNR). To choose a SMR value we apply again the $argmax$ criterion on $P(SMR|IRR, SNR)$:

$$\hat{smr} = \arg\max_{smr}(P(UA = smr|E = \{irr, snr\})). \tag{7.8}$$

7.8 Discussion on multimodal grounding

7.8.1 Efficiency of the repair strategy

Introducing two phases of grounding results in the advantage that we do not need to provide all the evidence from the input modalities in the first grounding phase when the robot is concerned with the issue of user presence and attention to the dialogue. Running a speech recognition process at this stage will just result in unnecessary work load for the robot system. On the other hand, the task of people detection and face detection are also required for the purpose of safe navigation and situation awareness of the mobile robot. They are typically implemented and running all the time and their status is already available. Thus, the two-phase separation of the grounding process contributes to the efficient utilization of the robot modality information. It also defines an efficient strategy for communication failure detection and repair. Given the dependencies in the Bayesian network in Figure 7.5 (a), inferring that $UA = 1$ is causally related with $UR = 1$. In other words, a presence of a face in the video stream would imply presence of legs in the laser scanner reading. Thus, the user presence (UR variable) is checked only when UA is inferred to be 0, using the $argmax$ criterion on the UA posterior probabilities. The two phases of grounding in our model allow an explicit schedule for tracking of the grounding state values. This schedule is shown in the form of a decision tree with four possible outcomes in Figure 7.6. Shaded nodes in the tree denote inference and $argmax(\cdot)$ evaluation of a particular grounding state (e.g. S0:UR). The tree is processed top-down (from the root - UA down to the leaves that represent the four possible outcomes in our case). The outcome 4 triggers the process of user goal identification from the interpreted recognition result (IRR) values. The remaining outcomes can be used to trigger modality-specific multimodal repairs.

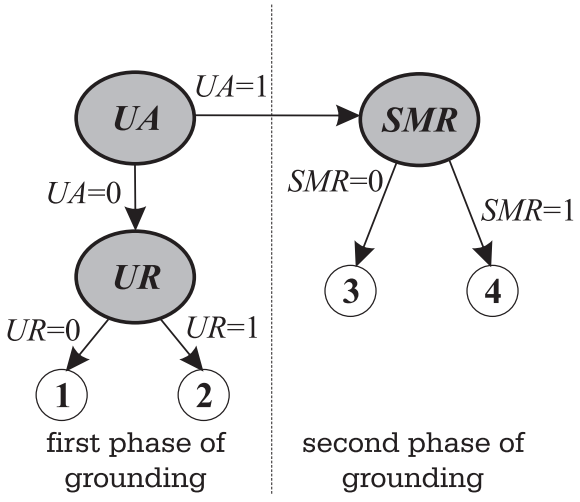


Figure 7.6: Decision tree for two phase grounding

7.8.2 Grounding with multimodal dialogue repairs

In order to consider a grounding state as being reached, the robot seeks for a probability above chance level for a particular value (e.g. $UR = 1$) of the modality event associated with that state given the evidence from its input modalities. Hence, we have established a grounding criterion for the purpose of service robot dialogue that is based on the probability of the modality events associated with the grounding states in a two phase grounding model. Whenever a failure to reach a state is detected the multimodal grounding model introduced in this section can be used to trigger multimodal dialogue repair techniques (grounding actions). For example, failure to reach grounding state S0 ($UR = 0$) can trigger a dialogue repair action dedicated to finding a user ("Search visitor"). This repair action can combine speech synthesis as well as the move modality of the robot in the process of user search. If the second state of grounding is not reached ($UA = 0$), speech as well as the robot expressive face can be used to attract the attention of the user. Buttons can be used as an alternative input when the grounding state S2 is not reached (SMR=0). At the end, if the user goal is still undefined (UG=0) the expressive face along with the speech synthesis can be used to hint the user for the possible keywords that her/his answer can contain. In order to model the robot preferences on a particular repair action the framework of decision networks and utilities presented in the first part of the chapter can be directly used with the presented model of grounding. The grounding model can be also applied in more complex dialogue systems employing keyword spotting as well as continuous recognition systems in a system-initiative or mixed-initiative dialogue setting. In particular, the first phase of grounding would not require any modification or changes in the network topology. As long as we preserve the user goal-oriented turn structure of the service dialogue, the second phase of grounding may not require any changes in the network topology either. We have to mention however that depending on the representation of the user goal (Hong et al., 2005) and the type of the recognition task involved (keywords, continuous speech), the grounding model in its second phase may need additional states associated with speech-based dialogue repair acts well known from the spoken dialogue literature (e.g different kinds of confirmation and disambiguation grounding acts (Brennan and Hulteen, 1995)).

7.8.3 Scalability of the grounding model

As outlined in Section 7.4, extending the model with additional modalities and user goals should be done after taking into account the complexity issues concerning the framework of Bayesian networks. The computational complexity of exact inference in Bayesian networks with conditional Gaussian pdfs is NP hard (Murphy, 2002; Cooper, 1990). In our case however, the use of two phases of grounding and special Bayesian network topologies lead to great reduction in the computational demands for inference in the Bayesian networks in Figure 7.5. In addition, the continuous variables are all observed, which avoids the problem of marginalizing continuous variables.

The Bayesian network in the first phase of grounding is a member of a special class of Bayesian network topologies: the polytree or the singly-connected networks, that allow linear dependence of the number of computations needed by exact inference on the size of the network (Chapter 4). A polytree network is a Bayesian network in which there is only one path between any two variables. The Baysian network in Figure 7.5 (a) is a polytree network that is a subtree from a more general topology depicted in Figure 7.7 (a). This network is composed out of slices corresponding to distinct modalities. Each such slice contains a modality event causally-related to a modality feature (Figure 7.7 (b)). The full topology in Figure 7.7 (a) can model the causal chains similar to the one in the first phase of grounding. For example, the modality event ME_2 can be the UA event (User is attending to the conversation) that in turn is seen as the cause for next modality event $ME_1 = UR$

(User is present staying in close range in front of the robot). In that case we end up with the network in Figure 7.5 (a). The incorporation of a new modality and its event/feature is straightforward - we just add a new slice in the causal chain. For example, ME_3 can be an event related the the event of a user who is speaking. This event can be seen as cause behind a feature MF_3 given by a voice activity detector. A slice can also represent another event from the same modality. For example the event of a speaking user can be related to a video modality feature related to detecting movements of the user's lips.

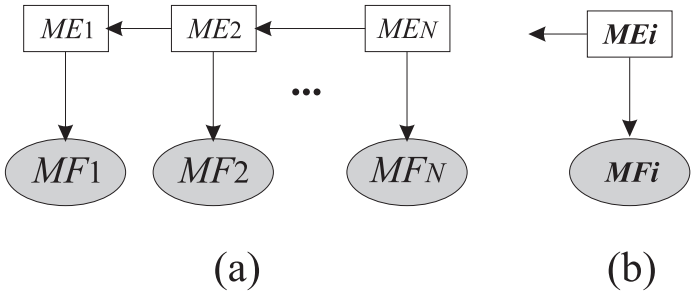


Figure 7.7: Bayesian network for grounding: (a) slice related to a modality event and its feature, (b) full topology

Using exact inference algorithms like the junction tree algorithm or variable elimination (Chapter 4; (Murphy (2002), Appendix B)) with the above BN topology will result in linear computational complexity $O(N)$ with the number N of the involved modalities events.

In the second grounding phase the definition of the SMR event also allows reduction in the number of operations needed by inference in the corresponding Bayesian network as already discussed in Section 7.7.2 (Equation 7.7). These observations demonstrate the important fact that particular Bayesian network topologies offered for multimodal grounding offer substantial reduction in the computational complexity of inference.

Modelling curiosity in the service robot behavior

The incorporation of new input modalities in the grounding model (Figure 7.7) results in introduction of new modality events and features. With the increased number of modality events it is more difficult to fully specify and motivate a systematic repair strategy as in the case presented in Figure 7.6. For example, consider the case of a first phase of grounding with three modality events, where we have three posterior values $P(ME_1|E)$,$P(ME_2|E)$,$P(ME_3|E)$. How can we decide which event from the set $\{ME_1, ME_2, ME_3\}$ should be evaluated first?

In answering the above question we can adopt one single idea, i.e. the idea of the service robot "curiosity". Let us assume that the service robot is interested in learning more about unknown than familiar facts. A fact for a service robot is a particular state of an input modality event. The more uncertain the state is, the more "curious" the robot will be about its state. In the field of information theory, the entropy is a widely accepted measure for quantifying the uncertainty of a random variable. The entropy of a discrete random variable is defined as follows:

$$H = -\sum_{i=1}^{N} P(X = i) \log(P(X = i)), \tag{7.9}$$

where X is a discrete random variable with N possible values, and $P(X = i)$ is the probability of $X = i$.

The entropy is maximized when the variable has a uniform distribution (maximum uncertainty) and is minimized when the variable is constant (there is no uncertainty in the variable's value). The entropy can be used for evaluating the uncertainty of about the modality events, i.e. $\{ME_1, ME_2, ME_3\}$. Then, a "curiosity"-driven service robot can choose to investigate the event with maximal uncertainty (maximal entropy). This modality event selection strategy can be helpful whenever the input modality events have equal importance in the grounding model. This simplified repair strategy could be a subject of future work.

7.9 Summary

In this chapter, we presented a complete methodological concept for designing and implementing repair strategies for avoiding communication failures in spoken dialogues with mobile tour-guide robots in mass exhibition conditions. In these conditions non-collaborative visitors' behavior and adverse acoustic conditions have been shown to be among the main factors for communication failures in speech-based interaction. The problem of tour-guide dialogue management is shown to depend on a robust inference of the user goal at each dialogue state, where the chance for communication failure can be explicitly modelled through an "undefined user goal". Bayesian networks are used to elicit probability distribution over the set of user goals, fusing acoustic (speech recognition result) and spatial (laser scanner signal) aspects of the user goal. In the case of high probability for the undefined user goal, dialogue repair sequences were chosen in accordance with the tour-guide requirements, exploiting different input and output robot modalities, e.g. speech or buttons-based input, move event, etc. Given that the real state of the user goal is never known for sure by the robot, the strategies for repair-action selection can be modelled using concepts from probability and decision theories and related graphical representations, e.g. Bayesian networks and their extensions - decision networks. Decision theory allowed us to define the tour-guide dialogue as a sequential process of decision-making, where decision networks were used to choose from the available actions at each dialogue state. Decision networks utilize a mathematical framework for choosing actions, based on the maximum expected utility (MEU) of the repair actions over the distribution of the user goals given by the Bayesian network. The MEU principle allows modelling of complex task-oriented tour-guide robot behaviors, through manipulating the utility function values.

In the chapter, decision networks were used for modelling tour-guide robot repair strategies, taking into account different aspects of the user goal. While the repair strategy, i.e. the sequence of repair actions, can be straightforward with two input modalities (e.g speech and laser), incorporating new modalities would require more systematic approach in designing time-consuming repairs. For this purpose we have introduced a multimodal state-based model for grounding conversation in the general case of service robots under noisy acoustic conditions. The model was motivated by reducing the risk of communication failures due to incorrect user goal identification with unprepared users in typical noisy robot deployment conditions. The model exploits the multiple modalities available in the service robot system to provide evidence for reaching grounding states. In order to handle the speech input as sufficiently grounded (correctly understood) by the robot, four proposed grounding states have to be reached. The initial two states are related to the events of presence of a user who is attending to the conversation with the robot. A Bayesian network combining information from the laser and video modality was used to estimate the probabilities that the grounding states have been reached. The remaining two states in the grounding model were related to the grounding state of reliable speech modality and the grounding state of valid user goal, i.e a user goal that can be mapped into a service provided by the robot. The speech modality reliability was explicitly modelled by the event of error in the user goal identification based on the observed recognition

result. Another Bayesian network was used to model the dependencies between the event of speech modality reliability, the user goal and the speech recognition result as well as the signal-domain measure related to the level of acoustic noise.

The criterion used to consider the conversation as grounded at each particular grounding state was based on the probability of the grounding state-related events, estimated by the Bayesian network. The use of two distinct phases of grounding has allowed us to utilize special topologies in the Bayesian networks that resulted in a reduced number of computations needed for the probabilistic inference. In particular, using a polytree (singly-connected) BN topology in the first grounding phase has allowed reduction from exponential to linear number of operations in the number of used modalities, needed by inference.

In order to test the performance of the model an evaluation protocol is needed. The type of dialogue used, based on recognition of keywords that can be mapped into user goals, suggests the use of accuracy in the process of user goal identification based on speech modality solely and when additional input modality information is used. These accuracies provide quantitative criteria for measuring the performance, however they are not sufficient when designing a communication interface to be used by people. In order to fully evaluate the usability aspect of the error handling techniques proposed in the thesis, combining the accuracy metrics with the results from user satisfaction tests are needed.

8 Experimental evaluation

This chapter is dedicated to the evaluation of methods proposed in Chapter 7 for error handling in human-robot dialogue using the multimodal grounding. The multimodal grounding utilizes probabilistic models built in the framework of graphical models, and is tested in a tour-guide dialogue scenario. In the chapter we focus on evaluating the benefit brought to the robot interactive system, by using a state-based grounding model for triggering multimodal dialogue repair actions after each user turn in dialogue. We demonstrate the evaluation methodology on the tour-guide service robot system RoboX. The evaluation is done on two levels, i.e. component and system levels, using technical (objective) and user-based (subjective) methods. On the component level the technical evaluation is done by using accuracies as objective measures of the performance of the grounding model and the resulting performance of the user goal identification after each user turn in dialogue. The benefit of the proposed error handling framework is demonstrated by comparing the accuracy of a baseline interactive system employing only speech recognition for user goal identification (Chapter 5) and a system equipped with a multimodal grounding architecture (Chapter 7). On the system level the technical evaluation is done with quantitative success criteria motivated by the tour-guide robot task requirements. Finally, results from subjective usability tests are compared with the results from the technical evaluation to assess the quantitative success criteria as system usability predictors.

Since the video modality was not recorded during Expo.02, we had to create new multimodal corpus to train and test the graphical models for the error handling system. Three different dialogue scenarios were designed during the multimodal data collection: 1) tutorial scenario for data collection, 2) simulation scenario needed by the technical part of the proposed evaluation, 3) normal tour scenario for user subjective test in baseline conditions.

8.1 Introduction

The evaluation of interactive systems is done in two main ways: technical and usability evaluation (Dybkjaer et al., 2004). The technical evaluation relies on the objective measures of system performance. For example, in the case of speech recognition system word accuracy (recognition rate) is a

widely accepted performance measure. Spoken interactive systems such as voice-enabled interfaces for robots employ multiple system components. One of them is the speech recognition system component. In the case of task-oriented dialogue, the system level evaluation metrics can account for dialogue task success, task completion time, turn correction ratio, etc (Dybkjaer et al., 2004).

On the other hand, interactive system usability evaluation remains very much dependent on the potential user. A system that exhibits good performance at the technical evaluation level are not guaranteed to be highly appreciated by their users. Users may prefer less-performant systems due to specific reasons like, for example, familiarity and associated ease of use, system price, etc. Therefore, usability evaluation remains largely subjective and user-oriented, relying on field tests and user feedback. User tests typically aim at assessing if the objective metrics used in system technical evaluation can be used as good predictors of the user satisfaction from the system.

The process of interactive system evaluation is initiated with a characterization step in which the particular system and components under evaluation are defined (Gibbon et al., 1997, 2000). The experimental evaluation presented in this chapter aims at assessing the effectiveness of a service robot dialogue system in achieving its objectives, when techniques for recognition error handling are introduced. The focus is on evaluating the impact of the error handling techniques using multimodal grounding (Chapter 7) on the effectiveness of human-robot interaction via voice. We perform the evaluation using the tour-guide robot RoboX. However, the presented evaluation methodology employs "black box" criteria and metrics that operate on the level of dialogue task success, and hence can be applied for every type of human-robot interactive system.

8.2 Interactive system characterization

The details concerning the interactive system used as baseline in the experiments (the dialogue system of the tour-guide robot RoboX) have been already discussed in Chapter 5. Therefore, we summarize the baseline system briefly, providing more details on the new interactive system that is equipped with a multimodal grounding.

In the case of the tour-guide robot RoboX, we have an interactive dialogue system in which the dialogue flow is guided by the system. The dialogue structure is represented as a sequence of exchanges (initiative/responce pairs) containing user and system dialogue turns in predefined order forming the complete dialogue scenario. The system dialogue turns end with a question addressed to the user. The recognition technique employed when the system is acquiring the user answer is based on word spotting with a small system vocabulary. The system questions have three answer alternatives: two words corresponding to two alternative user goals ($UG = \{1, 2\}$) and a third case of a undefined user goal ($UG = 0$) which can be expressed with every other word or combination of words.

The main task of the tour-guide dialogue system is to provide an interface for communication via voice between a human user and a robot providing guided tours. The robot has to attract a user and to guide the user through a predefined route, providing tour information (e.g. exhibit presentations) in an interactive fashion, ensuring that the user is following him and is attending carefully through the whole presentation.

8.2.1 Multimodal grounding model

The available input and output modalities on the robot platform are used in a process of multimodal grounding prior to identifying the user goals from the recognized words. The process of grounding is responsible for compensating for recognition errors that may arise due to the high noise level

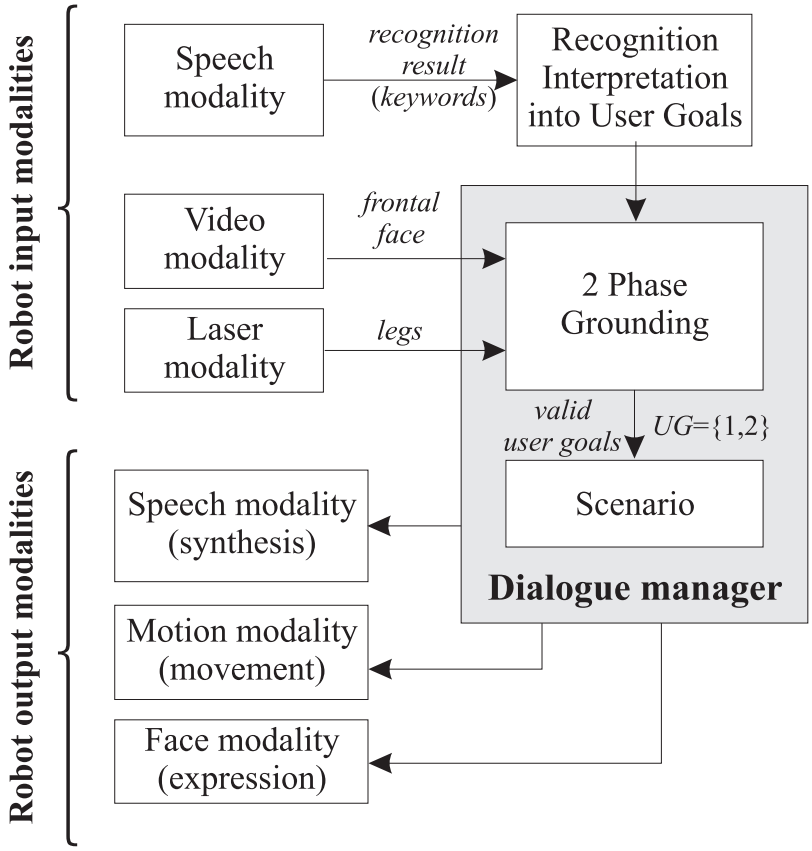


Figure 8.1: Tour-guide interactive system architecture.

or uncooperative user behavior at each dialogue exchange. The dialogue exchange is initiated by multimodal input acquisition and consequent system response through speech and other output modalities. The grounding process monitors the four grounding states and can trigger dedicated repair actions, depending on the grounding state values. The four grounding states correspond to the binary event of *U*ser presence in *R*ange for communication ($UR = 1$) as detected using the laser modality, the binary event of *U*ser *A*ttending to the conversation (looking in the robot's camera while speaking - $UA = 1$) as detected using the video modality, the binary event accounting for *S*peech *M*odality *R*eliability ($SMR = 1$) and the event of valid *U*ser *G*oal ($UG = 0$). The repair actions triggered when a grounding state is not reached (e.g. $UR = 0$ or $UA = 0$) manifest themselves as sub-dialogues that may employ other modalities along with speech. The main component responsible for the grounding process is the multimodal grounding model (Figure 7.5) that was described in details in Chapter 7. Figure 8.1 depicts how the grounding is fitted within the interactive dialogue system architecture of RoboX. Figure 8.2 depicts the repair dialogue used by RoboX with the help of the two phase grounding. The repair action dialogue sequences triggered by the grounding states are depicted in Figure 8.3.

The goal of the evaluation is to assess the impact of multimodal grounding on the accuracy of user goal identification on the system component level, and the influence of introducing grounding and multimodal repairs on the system efficiency on the system level, using objective technical evaluation metrics and subjective user-based methods for system usability evaluation.

In the stage of grounding, all four grounding states have to be reached to ensure that the user is

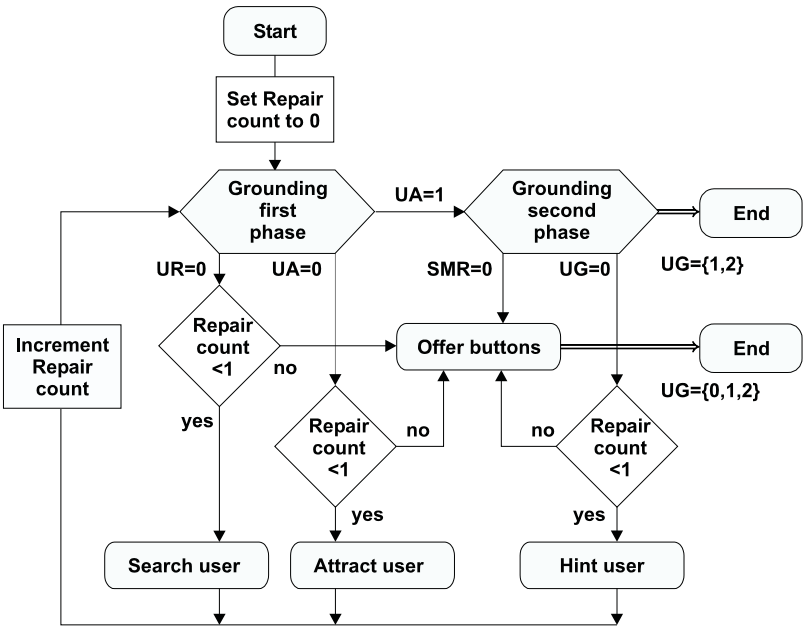


Figure 8.2: Tour-guide repair dialogue.

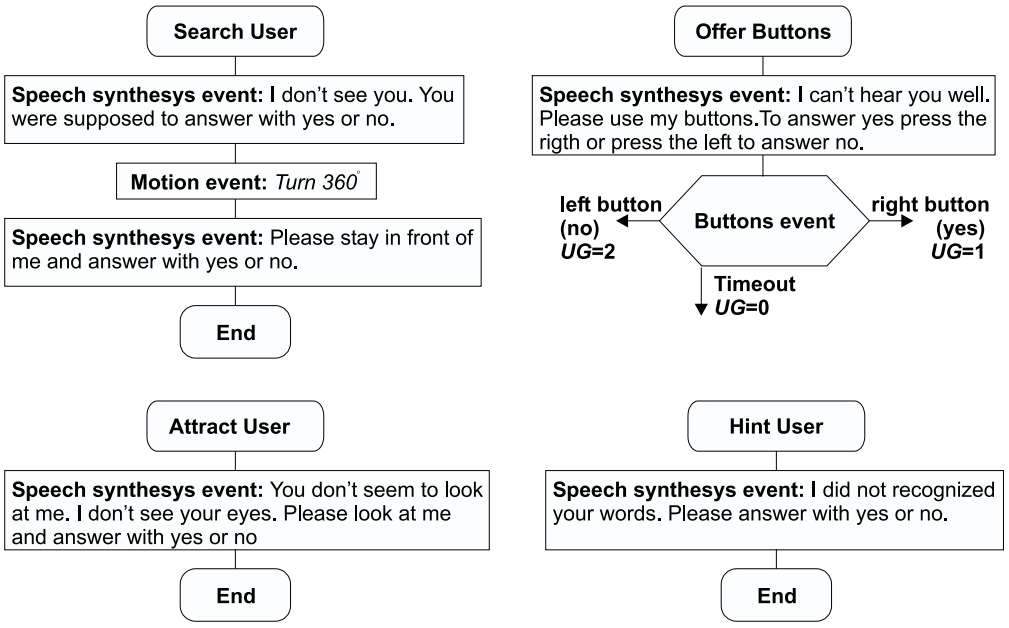


Figure 8.3: Tour-guide repair action dialogue sequences.

sufficiently attending to the conversation and understood by the robot (Chapter 7). The state-based grounding model is built using different modality data and graphical models for combining these data for inferring a probability on the possible grounding state values. To evaluate the technical aspect of the graphical model based grounding architecture, we have to provide sufficiently large multimodal corpus for training and testing of the corresponding graphical models.

8.3 Multimodal data set collection

8.3.1 The tour-guiding evaluation scenarios

In the evaluation experiments we use the mobile robot RoboX (Figure 7.5) as a tour-guide in the Autonomous System Laboratory at EPFL. In addition to the "yes/no" keywords in the initiative/response pairs, we have also used other keywords extending the possible user goals with services related to "general lab information" or "particular lab member location". The "general lab information" tour-guide service is in fact a guided-tour in which the robot moves along the corridors of the Autonomous Systems Lab, using yes/no questions to provide information about himself, the lab, the lab professors and their projects in interactive conversation (Table 8.1).

Robot:	So, do you want to have lab tour or you want to visit a particular person location. Say location or tour?
User:	Location
Robot:	I can show you the secretary's room. You can address location questions to her as well. I can also show you the room of my friend Plamen. So, who do you want to visit first the secretary or Plamen?
User:	Secretary
Robot:	Madame Marie Jo Pellaud is a very polite and helpful human being. Please follow me ... ...
Robot:	Are you already tired? Please answer with yes or no.
User:	Yes
Robot:	I will show you the coffee room after we finish. I hope you will enjoy the time spent with me. Please follow me now ...
Robot:	Look at the poster on the wall. Do you recognize me?
User:	No
Robot:	Well, you know I also can't recognize myself for the moment. May be, we both should ask for a software update. Do you want me to tell you a bit more about myself?
User:	Yes
Robot:	Thank you, I love to speak about myself. My purpose is ...

Table 8.1: Excerpts from the *normal* tour scenario

In the "particular lab member location" service the robot performs a guided-tour to the secretary's office or to the office of one of its developers. In order to distinguish between the above-specified services RoboX was configured to provide two new initiative/response pairs, i.e. "tour/location" and "secretary/plamen" in his dialogue turns. The user goal mapping was preserved as in the "yes/no" case, i.e. $UG = 1$ corresponds to $ORR = (yes|location|plamen)$, $UG = 2$ corresponds to $ORR = (no|tour|secretary)$ and $UG = 0$ to $ORR = GB$. The particular UG interpretation depends on the dialogue turns in the robot turn scenario sequence. The turn number was used for setting the particular speech recognition grammar as well.

We refer to the above dialogue scenario as the *normal* tour scenario. Our primary goal during the *normal* tour was to collect multimodal data for training and testing of the grounding model presented in Figure 7.5 as well as to observe the typical user behavior in order to identify the possible communication failures that our grounding model can address in the future. Since most of the time people were acting in a cooperative fashion during the tour, we have created a special *simulation* tour. The goal of this tour was to provide enough communication failure examples for the training

of the grounding model. During the *simulation* tour the robot itself was asking people to perform different behaviors corresponding to failures at the different states of the grounding model given in Table 7.3. In addition to simulate noisy conditions similar to the Expo.02 exhibition conditions, each turn was replicated and noisy audio files recorded from Expo.02 were played from the robot speakers during the data acquisition process. Excerpts from the *simulation* tour scenario are given in Table 8.2. A summary of the dialogue turns involved in the *simulation* tour scenario are given in Table 8.3.

Robot:	Hi, Nice to see you. The goal of this last session will be to record your "naughty" behavior. I will instruct you about tricks you should do to me. I hope you will enjoy this last part. So, go behind me and hide yourself while I am asking you a question. Press my buttons when ready and remember you are not supposed to answer my question. ...
User:	*A button is pressed*
Robot:	Do you want to have lab tour or you want to visit a particular person location. Say location or tour?
User:	...
Robot:	OK, now keep hiding. I will simulate background noise. Remember to stay behind me and to not answer to my question... ...
Robot:	OK, now the second trick. Stay in front of me but don't turn any attention to me. You can look aside or show me your back, you can also speak to people around. So, press my buttons when ready and remember you are not supposed to look in my eyes this time...

Table 8.2: Excerpts from the *simulation* tour scenario

Simulation scenario
Keyword vocabulary: yes, no, location, tour, plamen, secretary

Turn No	**Simulated failure**	**Description**
1	$UR = 0$	User absent
2	$UR = 0, SMR = 0$	User absent and noise
3	$UA = 0$	User not attending
4	$UA = 0, SMR = 0$	User not attending and noise
5	$UG = 0$	User remains silent
6	$UG = 0, SMR = 0$	User remains silent and noise
7	$UG = 0$	User utters out-of-vocabulary (OOV) words
8	$UG = 0, SMR = 0$	User utters OOV words and noise
9-14	$UG \neq 0, SMR = 0$	User utters each vocabulary keyword in noise

Table 8.3: Dialogue turn summary for the *simulation* tour scenario

To collect additional data for the training of the speech recognizer of RoboX and to make people familiar with the robot interface, we have also designed a *tutorial* scenario. In this scenario RoboX is explaining to people how to answer to it, asking them to repeat keywords from its recognition vocabulary five times.

Robot:	Hi, my name is Robox. I am the tour guide robot of the Autonomous Systems Lab. I hope you will enjoy the time spent with me. We will start with a Tutorial scenario. During the Tutorial you will learn how to interact with me and I will record what you say in order to improve my speech recognition later. Please be polite and look straight in my eyes when you talk to me. You can start speaking when my eye is blinking like now. Speak clear and loud as I'm sometimes a little deaf. So, let's start with some simple exercises. I will tell you a word and you will repeat after me. The first word is location. Now, be ready, look in my eyes and say location.
User:	Location
Robot:	Say location.
...	...

Table 8.4: An excerpt from the *tutorial* tour scenario

8.3.2 Data sufficiency issues

60 people were involved in the data set collection experiment (20 women and 40 men). The number of people was chosen according to the standard recommendations for minimal size, speaker-independent speech corpus (Gibbon et al., 1997). People were starting with the *tutorial* scenario, then they were asked to do the *normal* tour and the *simulation* tour.

During the *tutorial* scenario the new keywords to be recognized (*location*, *plamen*, *secretary* and *tour*) were repeated 5 times by each user. This particular number was chosen, based on the empirical recommendation that the number of training examples per recognized unit should be at least 5 times bigger than the number of the model parameters used in the recognition unit model. In the case of speech recognition, the phoneme is the basic building unit for each word. Phonemes in our recognition system are modelled with three state left-to-right hidden Markov models HMMs. We use four mixtures diagonal continuous Gaussian HMMs, in which the overall number of parameters per HMM state is equal to 14 (4 weights + 4 means + 4 variances + 2 transition probabilities). The phoneme HMMs are composed of three states, which results in 42 parameters in total. Following the empirical recommendations, if we assume that the phoneme inventory in our recognition vocabulary is uniquely represented in each vocabulary word, then the total number of training examples per word has to be at least 210 (5 times 42). We round this number to 200, since phonemes are repeated in some words (e.g. *location* and *plamen*). For testing purposes, we assigned two times less examples per word, i.e 100 in total. At the end the total number of needed examples per word becomes 300, which divided by 60 users resulted in 5 words per user for a given vocabulary entry. This number was a reasonable trade-off between the demands of training data and the time necessary for performing the needed data collection. The participating people were typically spending between 30-40 minutes communicating with the robot following the three dialogue scenarios (*tutorial*, *normal* and *simulation*). During these three dialogue scenarios, we collected data from four different input modalities of RoboX, i.e. laser, video, speech and buttons.

8.3.3 User detection

The laser modality was used for detection of the presence of a user in front of the robot ($UR = 1$ event, Figure 7.5). The scanners were located at a height of approximately 0.5 m, which makes it possible to detect the presence of the user's legs from the scanner reading. The leg pattern typically appears as two flat minima that resemble two lines in the 1D plot of the laser scanner reading

(Figure 8.4 (c)).

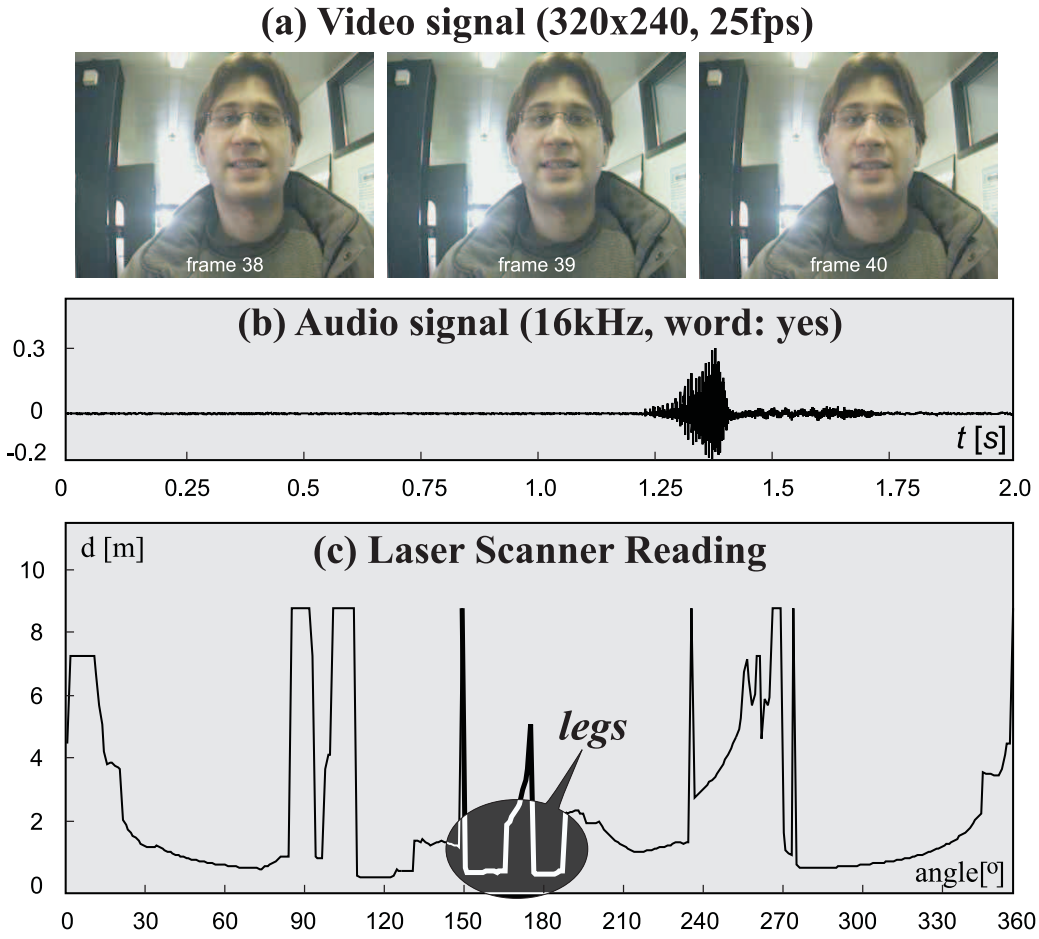


Figure 8.4: Video (a) Audio (b) and Laser (c) modality signals

Whenever the user is in range for communication (within 0.5 to 1.5m distance in front of the robot) the legs pattern typically appears as the closest object with respect to the the robot's front. Since we are interested in a possible user presence, the leg search is limited to the sector from the LSR (laser scanner reading) that corresponds to the robot's front. We have chosen an interval of 60^0 with respect to the robot front, i.e. the $[150^0, 210^0]$ from the LSR (Figure 8.4 (c)). The sector width is chosen to ensure that if the user is in front of the robot within the range for communication its legs are also in this sector. When the above condition holds the flat minima produced by the user's legs have a characteristic length of the flat parts. Since these flat regions are very similar to straight lines, the flat region length corresponds to the sum of the two lines lengths. Another interesting fact is that these two "lines" appear parallel to the x-axis into the 1D plot of the LSR. Since the robot is moving alongside a corridor such parallel patterns appear very rarely in the case of a missing user or they will be quite far from the robot. On the other hand, a histogram of the LSR produces high valued bins whenever such parallel structures are observed in the signal. The number of bins has to be chosen with respect to the needed precision when legs are detected. We chose 45 bins that divide the range of the SICK scanner into equally spaced intervals of 20 cm. In the case of a user present in front of the robot the first histogram bin is significantly higher compared to the case of no object, given that the robot is always looking alongside the corridor. Therefore, we

have chosen the first bin value for the continuous LSR variable used in by the Bayesian network in Figure 7.5 (a).

8.3.4 User face detection

The video modality was used for detecting a user attending to the conversation ($UA = 1$, Figure 7.5). Given the presence of a user, the robot has to detect if the user is attending to the conversation. We assume that presence of a user's frontal face in the video frames for an interval of time of at least 0.8 s is sufficient to ensure that the user is attending while providing her/his spoken answer. The video stream is providing 25 frames per second on the average (Figure 8.4). In order to provide evidence for the state of the UA variable from Figure 7.5 we use a face detector based on the modified algorithm of Viola and Jones (Viola and Jones, 2001; Lienhart and Maydt, 2002). To detect the user as attending we look for the binary event of face detected into 10 consecutive frames in the video stream. We assign this observed event a binary variable FD (face detected) and we use it in the Bayesian network in Figure 7.5 (a).

8.3.5 Speech modality reliability

The speech modality is used to obtain values for the observed variables in the Bayesian network in Figure 7.5 (b). The speech recognition system provides the values for the observed recognition result - ORR variable for each user turn in dialogue that are subsequently interpreted into IRR (interpreted recognition result into user goals) values. Each robot dialogue turn contains a question offered two possible services. The answer of the user is mapped into three possible user goals $UG = 1$ - first possible service, $UG = 2$ - second service and $UG = 0$ - undefined user goal at each dialogue state.

To measure the acoustical conditions affecting the noise factor (NF) we use a signal-to-noise ratio (SNR) related measure. The SNR can be defined as the ratio of the average energy of the speech signal divided by the average energy of the acoustic noise in dB. As in our case we have a single channel speech signal we estimate these energies based on two passes of audio signal acquisition. The first pass is just before the final question of RoboX and is 0.5 s long. The second pass is during the user answer and is limited to 2 s which was estimated to be a sufficient duration given the keyword vocabulary of RoboX. The signal n acquired in the first pass is associated with noise, while the signal s from the second pass is associated with speech. Our SNR-related modality quality measure (QM) is given by the formula:

$$QM = 10 \log_{10} \frac{\frac{1}{N} \sum_{i=1}^{N} s^2(i)}{\frac{1}{M} \sum_{i=1}^{M} n^2(i)}, \tag{8.1}$$

where $\{s(i)\}, i = 1, \ldots, N$ is the acquired speech signal containing N samples, and $\{n(i)\}, i = 1, \ldots, M$ is the acquired noise signal containing M samples. As the audio input of RoboX is sampled at $f_s = 16\ kHz$, then $N = 32000$, and $M = 8000$.

8.3.6 Database organization

The buttons modality of RoboX was used during the data collection to auto-assign user goals to the spoken answers of the user during the *normal* tour. In that case the users were asked to press one of the four buttons of RoboX corresponding to their spoken answer. The buttons status was recorded during the phase of input modality data acquisition, however the actual decision for the next robot dialogue turn was based solely on the speech recognition result (ORR) during the interaction with

the user. In the remaining two scenarios (*tutorial* and *simulation*) the user goals (UG values) were $a-priori$ known from the designing stage. The use of UG predefined scenarios (*tutorial* and *simulation*) and the buttons modality permitted automatic data tagging for all of the unobserved variables (UR, UA, NF, SMR, UG) in the robot grounding model.UG was set to 0, whenever UR or UA were 0. The NF values were set to 1 during the "noisy" turns in the *simulation* scenario (see Table 8.3) and 0 otherwise. According to its definition, SMR is 1 when UG coincides with IRR and is 0 otherwise.

8.4 Technical evaluation experiments

8.4.1 Component level evaluation

In the component level evaluation of the multimodal grounding we asses the accuracies of the grounding state predictors as well as the accuracy of the final user goal identification. The accuracies are calculated for the baseline tour-guide dialogue system and compared with an alternative system. The alternative system employs grounding and $argmax$ criteria on each of the grounding states posteriors to select a state value. It is named the "Argmax BN" system.

Accuracy computation

In the component level evaluation we adopt an accuracy metric similar to the word recognition accuracy as defined in the literature (Boros et al., 1996):

$$WAcc = 100\left(1 - \frac{N_S + N_I + N_D}{N}\right), \tag{8.2}$$

where N_S is the number of substitution, N_I the insertion and N_D the deletion errors. This measure is defined in general for utterances, where some words can be skipped (deleted) others can be inserted or substituted. In our case, the recognition task is to detect a keyword (e.g. yes, no, location, etc.) or a "garbage" word (GB) in the spoken input. Each keyword is distinct and directly mapped to a valid user goal (e.g. "yes" to $UG = 1$, "no" to $UG = 2$). The GB word is mapped to the undefined user goal ($UG = 0$). Therefore, the errors can be only of substitution type and we can directly evaluate the user goal accuracy using the formula:

$$Acc = 100\left(1 - \frac{N_S}{N}\right). \tag{8.3}$$

The same formula is used in the case of evaluating the grounding state prediction accuracy.

8.4.2 Accuracy of the "Argmax BN" system vs baseline system

The collected data set was used to train and test the grounding model networks. The full data set was used for training and testing of the attendance phase Bayesian network. Given the two phase grounding model of RoboX, the speech reliability Bayesian network was used only after detecting the event $UA = 1$ (User Attending) in the first phase of grounding. Hence, in the training of the second phase network, we do not really need data from the records for which UA is zero. Such data will very rarely appear in the second phase of grounding. For that reason the speech reliability phase Bayesian network was trained and tested on a partition of the full data set containing "clean" recordings ($NF = 0$) from the *tutorial* scenario and "noisy" ones ($NF = 1$) from the *simulation* scenario.

To test the accuracies of the individual grounding state predictor variables UR, UA and SMR we have run 50 cross-validation tests. Training and testing portions were chosen from the full and the partitioned data set each time at random. The size of the training portion was two times bigger than the testing portion. Values for the posteriors $P(UR|E_1)$, $P(UA|E_1)$ from the Attendance BN (Figure 7.5) and $P(SMR|E_2)$ from the speech Reliability phase BN were calculated for each testing sample ($E_1 = \{LSR, FD\}$ in the first case and $E_2 = \{IRR, SNR\}$ in the second case).

The values for the corresponding state predictor variables were assigned using the $argmax$ criteria (Equation 4.34) on the corresponding posterior probabilities. The tests were done for the events $UR = 1$, $UA = 1$, $SMR = 1$ computing corresponding accuracies. We have also done the tests for the noise factor event, i.e. $NF = 1$. The accuracies are calculated as the number of correct classifications minus the number of substitutions divided by the number of examples per class. The total number of training and testing examples were 1900 and 949 for the first phase of grounding and 1404 and 701 for the second phase. The accuracy statistics are given in Table 8.5.

Attendance BN Acc stats with 1900/949 train/test samples

Acc UR %:	$UR = 1$	$UR = 0$	**Total Acc**
μ	98.1	100	99.1
σ	0.3	0	0.3
Acc UA %:	$UA = 1$	$UA = 0$	**Total Acc**
μ	94.3	90.7	94.0
σ	0.6	3.2	0.6

Reliability BN Acc stats with 1404/701 train/test samples

Acc SMR %:	$SMR = 1$	$SMR = 0$	**Total Acc**
μ	89.9	67.6	83.5
σ	0.9	2.8	1.1
Acc NF %:	$NF = 1$	$NF = 0$	**Total Acc**
μ	80.6	93.5	90.6
σ	3.1	0.9	0.8

Table 8.5: 50 cross validation accuracy statistics for user attendance and speech reliability grounding phase BN models

To test the efficiency of the two phase grounding model in detecting the recognition errors, we have done the following experiment: We have trained the Bayesian networks in Figure 7.5 with a single iteration of the cross-validation test. The testing examples were provided first to the Bayesian network for the first grounding phase. If $UA = 0$ was calculated to hold after applying the $argmax$ criterion the user goal was set to $UG = 0$. Otherwise, the examples were provided to the second grounding phase Bayesian network. After computing the posterior distribution $P(SMR|E_2)$, if $SMR = 1$ was true, the IRR result (the user goal based on the speech recognition only) was used to assign a user goal. Otherwise, if $SMR = 0$ was selected after applying the $argmax$ criterion, we were setting the UG to its tagged value from the testing data. We assume that if the speech modality is unreliable and the user is requested to use the buttons the user goal is normally provided without errors.

The accuracy of IRR (the user goal based on the speech recognition only) was calculated and compared with that of UG after applying the two grounding phases. The results are presented in Table 8.6.

As can be seen from Table 8.5, the grounding state predictors function significantly above chance level. Thus, should the grounding level need to be assessed, the cause of the communication failure

can be located and remedied. This statement seems to be strongly supported by the results from our evaluation experiment as well. As can be seen from Table 8.6 the use of the Bayesian networks in Figure 7.5 for the two phases of grounding has resulted in a significant improvement in the accuracy of the user goal identification. The gain in performance is due to the improved identification of the garbage case $UG = 0$, which in turn is due to the good detection rate of the event $UA = 1$ in the first phase of grounding when using the Bayesian network in Figure 7.5 (a). Modelling of the event of error in user goal identification based only on the observed speech recognition results in the second phase of grounding and the availability of an alternative input modality (interactive buttons) can enable even further improvement in the user goal identification as demonstrated in Table 8.6.

Total Acc IRR	$IRR = 0$	$IRR = 1$	$IRR = 2$
67.13 %	63.33 %	65.25 %	72.81 %
Total Acc UG	$UG = 0$	$UG = 1$	$UG = 2$
90.21 %	95.00 %	84.40 %	91.23 %

Table 8.6: Statistics about user goal identification before (IRR) and after grounding (UG) on 315 testing examples

8.4.3 System-level evaluation

Mobile service robots in general and tour-guide robots in particular are physical agents that act in the real world, sensing changes in the environment through their input modalities (e.g. speech) and performing actions through the output modalities (e.g. synthesized speech). The performed action at each time, given the information acquired from the input modalities at that time has to be chosen in order to maximize a performance metric. The performance metrics are measurable quantities related to success criteria that evaluate how successful the agent is in fulfilling its communicative tasks. Given the discussion provided in Chapter 5, and 7 on the tour-guide task requirements, we can adopt the following criteria for evaluating a successful tour-guide dialogue.

A tour guide robot is considered successful in its interaction with the user if:

Criterion 1. The user is attending to the conversation, which means that the states in the first phase of grounding are reached in all initiative/response pairs, during one full tour-guide dialogue scenario. In this way, we ensure that information is successfully conveyed to the user.

Criterion 2. The user choice is considered after each user turn in dialogue. In other words, user goals are correctly identified during the dialogue. This additionally requires that the states in the second phase of grounding are reached in all initiative/response pairs, during dialogue.

The above success criteria can be related to the well known ”black box” evaluation metrics used for the technical evaluation of dialogue systems. For example, Criterion 1 can be related to the dialogue task success metric. In the case of tour-guiding, completing a full scenario with a user attending to the conversation can be seen as a successful task completion. We assume that the fact that the user is following the robot and attending to the conversation clearly signifies that the user is interested.

The criteria are ordered according to their decreasing significance for the usability perspective of the voice-enabled tour-guide robot. If a user is always present and attending (Criterion 1) in front of the robot, we assume that the level of user interest and interface usability is high. Although user goal identification accuracy is important from the perspective of the tour-guide ability to provide

desired service it is not assumed to be more important than the ability of the the tour-guide to attract its users. The final goal of providing specific information should not contradict the goal of keeping the user involved and informed according to her/his intent.

Criterion 1 can be quantified by the parameter "user attendance rate". We define it as equal to the number of times during the dialogue that the user was attending to the conversation ($UR = 1$ and $UA = 1$) divided by the total number of robot dialogue turns:

$$UAR = \frac{1}{N} \sum_{t=1}^{N} I_t(UA = 1), \tag{8.4}$$

where UAR is the user attendance rate and $I_t(UA = 1)$ is the indicator function of the event $UA = 1$ at each dialogue turn $t = \{1, ..., N\}$. The number of dialogue turns N in the definition does not include the additional repair turns. In order to have a "fair" measure, the indicator function has to be used with *a priori* annotated reference state after looking at the collected dataset. It can be also "unfair" if we take the automatic state value as given by the grounding model after performing the dialogue repair sequence (Figure 8.2).

Criterion 2 can be directly quantified by the user goal identification accuracy during the spoken interaction. In addition, to evaluate the efficiency of considering the user choice using the grounding model for multimodal dialogue repair we introduce the so-called *Repair proportion*. The repair proportion is closely related to the reported turn repair ration metric in dialogue system evaluation (Dybkjaer et al., 2004). The *Repair proportion* is calculated with respect to the number of robot dialogue turns in the dialogue, i.e.:

$$RP = \frac{N_{repairs}}{N}, \tag{8.5}$$

where RP denotes the repair proportion measure, $N_{repairs}$ corresponds to the total number of repair turns, and N corresponds to the number of dialogue turns in the current dialogue scenario as in Equation 8.4.

All the metrics specified above had to be calculated for the baseline dialogue system and after performing grounding and corresponding repairs to evaluate the yield from applying the error repair techniques, using the *normal* tour described in Section 8.3.1. However, the data collected with the *normal* tour for the purpose of the component-level technical evaluation (Section 8.3) were recorded under controlled user conditions. In order to get the real figures using the above system-level metrics, we need an interactive scenario that is close to the real conditions of application. For that purpose we have performed a subjective user satisfaction test, where the system-level objective metrics are calculated and compared with results from user surveys on the interactive system usability.

8.5 Subjective user satisfaction tests

In the subjective user test 22 users (7 female / 15 male) are asked to perform the *normal* tour scenario. There were not given any additional information apart from a very general description of the robot and its input modalities. In addition, the tour itself is initiated with a short help on how to communicate with the robot. During the tour the user was advised to behave as natural as possible. The user was not obliged to follow the whole presentation if she/he gets very bored or for any other reason was willing to abandon the robot. Table 8.7 depicts statistics about the people involved in the experiment.

The main focus of the experiment was on the ability of the robot to keep its user involved and attending to the interaction. At the end the user was given to fill in a survey that aims at assessing

User No	Occupation	Sex	Age between:		English speaker	Familiarity with Robots*		
						1	2	3
1	PhD student	female	25	35	non	no	no	no
2	Student	female	25	35	non	no	no	no
3	Student	female	25	35	non	no	no	no
4	Unemployed	male	36	45	non	no	no	yes
5	PhD student	male	25	35	non	yes	yes	yes
6	Assistant	male	25	35	non	yes	no	yes
7	PhD student	male	25	35	non	no	no	no
8	PhD student	male	18	24	non	no	no	no
9	PhD student	male	18	24	non	yes	no	no
10	Post-doc	male	36	45	non	no	no	yes
11	Professor	male	25	35	non	no	no	yes
12	PhD student	female	25	35	non	yes	yes	no
13	Assistant	male	25	35	non	yes	yes	yes
14	PhD student	male	25	35	non	no	no	no
15	PhD student	female	25	35	non	no	no	no
16	PhD student	male	25	35	non	no	no	no
17	PhD student	male	25	35	non	no	no	no
18	PhD student	male	25	35	non	no	no	yes
19	PhD student	male	25	35	non	no	yes	yes
20	PhD student	female	25	35	non	no	no	no
21	Musician	female	25	35	non	no	no	no
22	Scientist	male	46	55	non	yes	yes	yes
Average:	**PhD student**	**68%**	**26**	**36**	**100%**	**73%**	**77%**	**59%**
Comment:	**mostly**	**male**	**-**	**-**	**non**	**no**	**no**	**no**

* 1. Have you ever used a real robot?; 2. Controlled a robot with voice?; 3. Used speech recognition software?

Table 8.7: Personal information about the user satisfaction test participants

the user satisfaction with the interactive performance of the RoboX system. The survey questions can be found in Appendix C.1.

During the user satisfaction test the multimodal user input (speech/video/laser) was recorded along with the status of the repair dialogue sequence. This status includes the number and type of the performed repairs during the repair dialogue sequence, including the detected grounding state value. At the end of each *normal* tour the real grounding state values manually annotated are compared with the automatically detected ones and system-level evaluation metrics are calculated. To evaluate the gain from the use of repair actions, the system-level evaluation metrics are calculated before and after the repair sequence. The results after extracting information from the user survey and calculating the system-level evaluation metrics are presented in Table 8.8. The gains in user goal identification and the user attendance rate before and after repair are visually depicted by the histograms in Figures 8.5 and 8.6.

The two subjective measures of system usability presented in Table 8.8 (**Dialogue quality** and **Recognition performance**) were extracted from the user answers to questions 1 and 7 in the survey. These questions along with the answers statistics from the 22 participants are depicted in Figure 8.7. The user satisfaction with the repair sequence performance is depicted similarly in Figure 8.8.

Finally the results from the user satisfaction tests were compared and correlated with the calculations of the three metrics for system level evaluation (user attendance rate, user goal accuracy, and repair proportion - Section 8.4.3). The results from the comparisons are depicted in Table 8.9. The correlation between the metrics is evaluated by calculating the correlation coefficient. The correlation coefficient ranges from -1 to 1, where an absolute value of the coefficient over 0.5 indicates strong correlation between the corresponding metrics.

	Subjective metrics:		Technical metrics:					
UserId	**Dialogue Quality**	**Recognition Performance**	**Task Success**	**Repair proportion**	**UG Acc** before repair**	**UG Acc after repair**	**UAR* before repair**	**UAR after repair**
1	5	2	0	1.11	0.56	1.00	0.67	0.89
2	4	4	1	1.23	0.31	0.92	0.85	0.92
3	5	5	1	0.30	0.80	0.80	1.00	1.00
4	3	3	1	0.69	0.71	0.86	1.00	1.00
5	5	4	1	0.40	0.70	0.90	0.90	1.00
6	3	2	1	1.00	0.44	1.00	0.78	0.78
7	3	3	1	0.78	0.33	0.89	0.67	1.00
8	4	3	1	0.56	0.80	1.00	0.90	0.90
9	3	3	1	0.33	0.64	0.82	1.00	1.00
10	3	3	1	0.46	0.75	1.00	0.85	0.92
11	4	4	1	0.33	0.78	1.00	1.00	1.00
12	4	4	1	0.56	0.56	1.00	0.89	1.00
13	4	3	1	0.46	0.69	0.92	0.92	1.00
14	5	5	1	0.33	0.89	1.00	0.89	1.00
15	4	4	1	0.89	0.44	1.00	0.89	0.89
16	3	4	1	0.56	0.78	0.89	1.00	1.00
17	4	4	1	0.55	0.55	1.00	1.00	1.00
18	4	4	1	0.33	0.78	1.00	0.89	1.00
19	3	1	1	0.56	0.67	1.00	0.89	0.89
20	5	4	1	0.44	0.78	1.00	1.00	1.00
21	2	2	0	1.19	0.67	0.76	0.62	0.67
22	4	3	1	0.54	0.62	0.92	0.92	1.00
Average:	**3.82**	**3.36**	**0.91**	**0.62**	**0.65**	**0.94**	**0.89**	**0.95**

Comments: * UAR - User Attendance Rate; ** UG Acc - User Goal Accuracy.

Table 8.8: Comparison between subjective user satisfaction and system level evaluation metrics

		Subjective metrics:	
	Correlation coefficient	Dialogue Quality	Recognition Performance
Technical metrics: **UG Acc** - User Goal Accuracy; **UAR** - User Attend-ance Rate	Repair proportion	-0.3273	-0.4593
	UG Acc before repair	0.2341	0.2617
	UG Acc after repair	0.3491	0.0186
	UAR before repair	0.2852	0.5215
	UAR after repair	0.4856	0.6064

Table 8.9: Correlation between the subjective user satisfaction and the technical system-level evaluation metrics

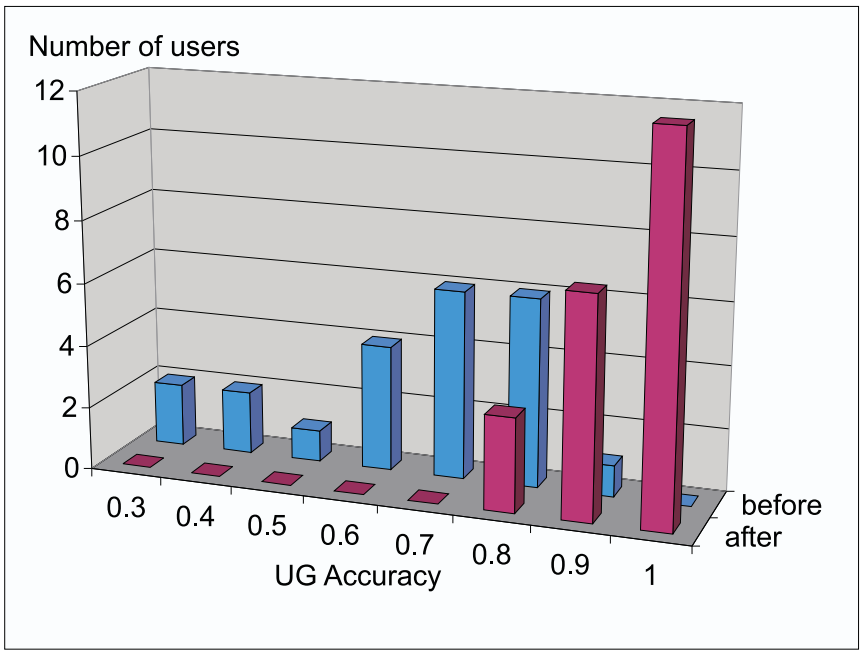


Figure 8.5: A histogram of the accuracy of User Goal (UG) identification before and after performing repair actions

8.6 Discussion

8.6.1 System-level evaluation metrics and system usability

Despite the limited available data (22 participants) the results presented in Table 8.9 can still provide intuitive interpretations. The correlation statistics show that the subjective measures of system usability correlate most strongly with the attendance rate of the user and less strongly with the user goal identification accuracy before and after executing any repair actions. In other words the interested and attending users tend to appreciate the dialogue scenario with RoboX, and give good feedback for the system, supporting the assumption behind Criterion 1 from Section 8.4.3.

The repair sequence duration also illustrates an ”intuitive” correlation with the user appreciation of the system. Non-surprisingly, users seem to be less tolerant to frequent repairs. As mentioned by some of them, frequent repairs could give the user the feeling that the system is not operating properly although the high accuracy of the final user goal identification has been achieved.

The grounding state detection accuracy and subsequent repair action selection was also highly appreciated by the users as demonstrated by the results in Figure 8.8. The use of multimodal repairs was essential for the remedy of the wrongly assigned user goals, although the initial user goal identification accuracy does not show strong correlation with the subjective user satisfaction measures in the considered data pool.

Finally, both technical and user-based evaluation supported the fact that the proposed grounding architecture can contribute to a significant gain in the accuracy of final (after the repairs if any) user goal identification (Table 8.6, Figure 8.5), as well as a gain in user attendance rate (Figure 8.6). Hence, the use of multimodal grounding can enhance the usability of the service robot interactive system. The above statement is also supported by the high average task success and UAR (user attendance rate) with the 22 users (Table 8.8). It has to be taken into account that in real application conditions users may be less cooperative than the participants in the presented user study.

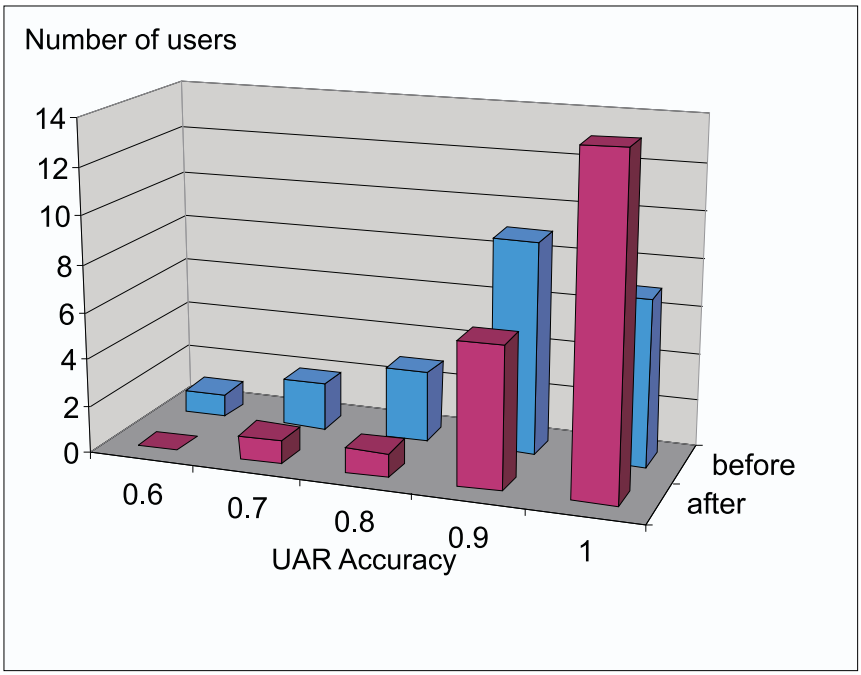


Figure 8.6: A histogram of User Attendance Rate before and after applying repair actions

According to the two subjective usability measures and the technical measures from Table 8.9, we can conclude that the RoboX dialogue scenario was appealing to the user and that the robot was efficient in providing its information to its user. In order to provide finer interpretation and motivation behind the above statement in the following section we perform a communication failure analysis of the logged grounding state values during the user tests. The user feedback is also analyzed to provide guidelines for further improvement of the interactive system of RoboX.

8.6.2 Communication failure analysis

During the user tests there were two cases in which the interaction between RoboX and its user has resulted in a communication failure (the robot was unable to identify a valid user goal after two consecutive repair actions). In the first case, the user wanted to experiment with the robot on purpose, and did not answer the robot's questions to see what will happen. After the buttons repair timed out RoboX left, informing that if the user is still there they can meet again near the coffee room. The second case was due to technical problems with the video camera. As a result the user was repeatedly asked to look the robot in the eyes without a real reason for such a repair action during several consecutive system turns. The increased repair activity frustrated the user, who finally left the robot to look for a human operator. As a result RoboX moved to the coffee room area, where after re-plugging the camera cable, the robot operated without any further technical problems.

Among the main sources for errors in user goal identification when only speech recognition was used, were the background noise, particular user accents or clipping of the user answer, because of the two seconds acquisition time interval. In such conditions the subsequent repairs were useful giving the robot a second chance for input acquisition, as well as the alternative to use buttons in the case of noise and in the second repair pass. Due to the "two phase" SNR calculation technique described in Section 8.3.5, whenever the user answer was preceded by non-stationary (temporary) burst of

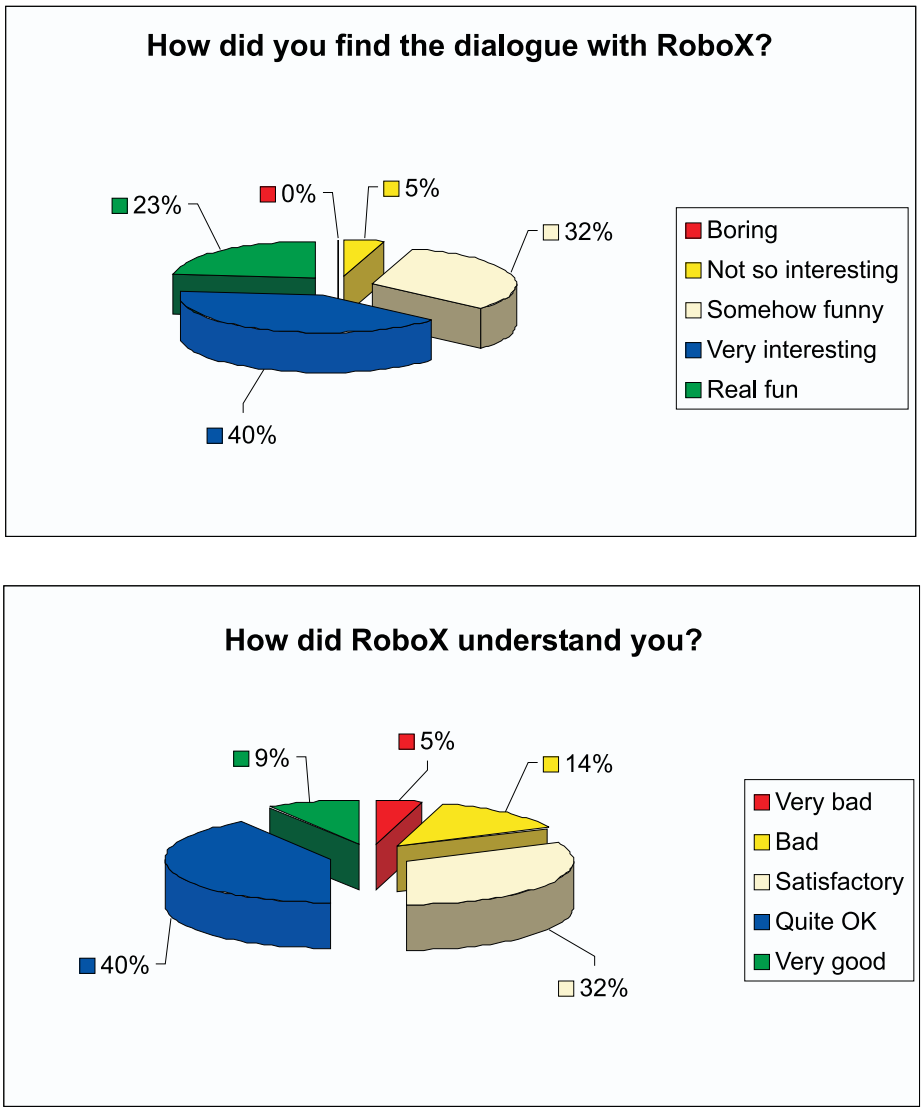


Figure 8.7: User satisfaction with the dialogue quality and the recognition performance during dialogue

noise, the robot was declaring the user answer as very noisy, although it was actually recorded in clean audio conditions. A particular case of errors resulted from low frequency audio modulations, caused by a mechanical vibration affecting the microphone, producing high energy values for the noise segment in comparison with the speech energy during the user answer. Users could produce such vibrations by applying stronger push on the interactive buttons prior to answering verbally to the robot. Disturbances in the quality of the acquired audio signal, unrelated to the acoustic environment, were additionally observed when the robot was operating near electrical facilities, such as electric boxes and transformers in the lab corridor. Such audio disturbances could potentially result in wrong repair actions related to speech modality reliability.

Detecting the state of user attendance depends directly on the frontal face detection accuracy. With proper user positioning with respect to the camera the errors in face detection were mainly due to adverse illumination conditions, i.e. sun flare from behind the user during the day or insufficient

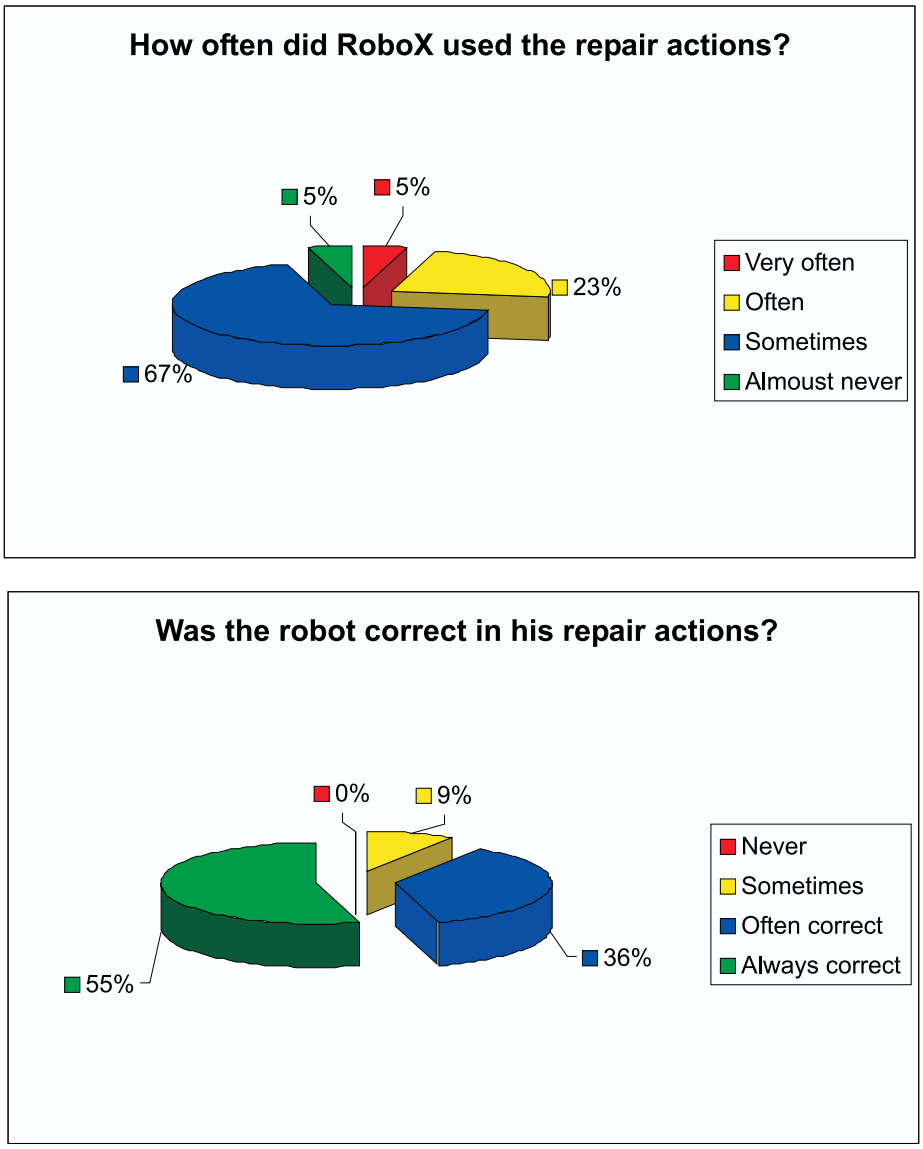


Figure 8.8: User satisfaction with the dialogue repair quality

light in the evenings. The other main source of errors resulted from the user posture or camera adjustment. In these cases, typically, part of the face was remaining outside the visual range of the camera. This was often the case with users that tended to stay too close to the camera or tended to bend towards the microphone while answering. Clipped faces also resulted with the users that were staying aside instead of directly facing the robot's front. In the last case, the "Attract user" repair was particularly useful for successful grounding (reaching the state of UA=1).

Finally, in some repair sequences users pressed a wrong button that resulted in a wrong user goal assignment. Nevertheless, in general, users remained interested in the conversation. Sometimes, incorrect user goals remained even unnoticed or were attributed to the humoristic character of the robot.

8.6.3 User feedback

As seen from Figure 8.7 most of the user test participants described the interaction with RoboX as funny and entertaining. Since many of them were unfamiliar with robots (Table 8.7) and with the dialogue scenario, the system driven dialogue did not make a bad impression on them. There were no recommendations in the survey that explicitly suggested changing the dialogue initiative. However, several persons recommended the robot to use more keywords and be more personal with them (e.g. asking for their name and using this information in the scenario). One of the users even started spontaneously answering with natural speech, but after the second question he understood that the robot preferred keywords, and adjusted his spoken answers appropriately.

People found the humoristic style of the tour guide as appropriate for its task. When asked if they would prefer "more serious" tour-guide, all users answered negatively. The positive attitude towards communicating with the robot did not change even when the robot's speech recognizer was not performing well all the time. However, in these cases the repair style was found to be important in order to avoid the impression that the system does not perform well. One user that exhibited low recognition performance (numerous "Hint user" repairs in more than two consecutive dialogue turns) recommended that the input modality should be permanently switched to buttons after given number of repairs related to speech recognition. Another user perceived the repeating "Attract user" repair as impolite, suggesting that the repair text should vary to overcome this impression. In two of the cases in the study with high repair activity (Repair proportion > 1), the users reported that their high concentration in answering the robot has distracted them from the normal process of listening to the information content provided by the robot during the tour-guiding scenario. However, most of the users (86 % - 19 out of 22 people) reported that the repair actions helped them stay involved and more interested in the dialogue. The repair actions seemed to distract people from their sometimes "destructive" desire to investigate and experiment with how they can put the robot in difficulty. We have to mention however that users were mostly highly educated people aware of the fact that the robot is recording their activities. Throughout the scenario the user preference towards the two alternative input modalities remained mostly in favor of speech and the combined use of speech and buttons. Only two of the users preferred permanently the use of buttons.

The complete user satisfaction tests survey results are presented in Appendix C.2.

8.6.4 On the use of alternative modalities

The availability of an alternative input modality with higher reliability than speech recognition in identifying user goals proved to be important for the process of service dialogue error handling. Buttons were appreciated by people and were a preferred modality in the cases when speech recognition did not work due to background noise. The buttons modality also gave a possibility of detecting clearly a communication failure in the case of buttons' timeout during input acquisition. After the buttons' timeouts the grounding status log could be investigated for the particular failing grounding state.

8.7 Summary

In this chapter we have presented an evaluation framework for a multimodal grounding architecture for triggering repair actions in spoken interaction between a human user and a mobile tour-guide robot.

The evaluation was done in two parts: technical evaluation and user-based subjective evaluation. Technical evaluation was based on measures of accuracy on the component level of grounding-state

prediction as well as global measures of dialogue task success. We introduced two global measures for tour-guide dialogue success, i.e. task success and user attendance rate.

In the user-based evaluation we have conducted an experiment with 22 users that were using the dialogue with grounding. We investigate the correlation between subjective user satisfaction derived from a user survey with the dialogue success metrics. Despite the limited available data (22 participants) the results provide intuitive interpretations. The correlation statistics show that the subjective measures of system usability correlate more strongly with the attendance rate of the user and less strongly with the user goal identification accuracy before and after executing any repair actions.

The use of the multimodal repairs was essential for the remedy of the wrongly assigned user goals, although the initial user goal identification accuracy does not show strong correlation with the subjective user satisfaction measures in the current data pool.

Finally, both technical and subjective user satisfaction evaluation supported the fact that the proposed grounding architecture can contribute to a significant gain in the accuracy of the final user goal identification, as well as a gain in user attendance rate. The evaluation shows that generally, the use of multimodal grounding enhances the usability of the service robot voice-enabled communication interfaces.

9 Conclusions

In this thesis, we have developed new methods for speech recognition integration in an interactive voice-enabled interface of a service robot, in particular a tour-guide robot. The methods exploit the use of probabilistic graphical models in two main directions: the modelling of probabilistic fusion of different input modality information for preventing and detecting communication failures in human-robot spoken interaction, and the use of multimodal strategies for communication failure recovery based on combining speech and other input and output robot modalities.

The main contributions related to probabilistic modality fusion using graphical models include: the formulation of the problem of multimodal user goal identification in tour-guide dialogue in the probabilistic framework of Bayesian networks, the introduction of error handling models based on low-level grounding between the tour-guide robot and its user in the spoken interaction, and modelling grounding in human-robot interaction using Bayesian networks tailored to limited computational resources.

The contributions related to the use of multimodal strategies for human-robot communication failure repair include the development of dialogue repair methods based on the use of dialogue repair sequences that exploit different robot modalities. We introduce principles from decision theory and related graphical models in the repair strategy to select the most appropriate multimodal repair action in the light of the tour-guide task requirements. In the design and execution of the repair strategy in human-robot spoken interaction, we contribute by providing a systematic approach based on the use of a state-based grounding model using Bayesian networks fusing information from speech laser and video robot modalities.

Mobile service robots are going to play an increasing role in the society of humans. Voice-enabled interaction becomes very important, if such robots are to be deployed in real-world environments and accepted by the vast majority of potential human users. Although an input modality based on speech recognition can exhibit degraded recognition performance during human-robot interaction in real-world noisy environments, we have demonstrated in the thesis that speech recognition can be successfully used in the human-robot interaction, when information from other input modalities such as laser and video is available. In this way, the work presented in the thesis is an important contribution to the emerging field of human-robot spoken interaction with multimodal voice-enabled

interfaces. In the remainder of this chapter we make a detailed review of the thesis contributions, outlining open issues for future research.

9.1 Modality fusion for error handling in communication with tour-guide robots

The main requirement for a tour-guide dialogue system is to present as much as possible exhibit information to the interacting visitor (user) in a limited time. The short-term human-robot interaction in mass exhibition conditions, where visitors and robots produce high level of acoustic noise motivate a robot-driven dialogue flow, relying on keywords or short meaningful phrases.

Our field experiments with RoboX during the Swiss National Exhibition Expo.02 showed that such an interaction scheme could be seriously challenged by the visitors' behaviors and the adverse acoustic conditions, causing errors in the speech recognition. Standard techniques for error handling in speech recognition are based on error detection and correction and usually use recovery dialogues. However, detecting errors using only speech recognition can be difficult and repair dialogues may be inefficient in the acoustic conditions of mass exhibition.

We introduced a new approach for error handling in spoken dialogue systems for mobile tour-guide robots working in mass exhibition conditions. The framework of Bayesian networks was introduced for detecting and correcting errors in the user goal identification in human-robot dialogue using multimodal input. We demonstrated that a Bayesian network can model efficiently the dependencies between the speech and the laser scanner input modality information. In addition, we modelled explicitly the speech recognition reliability, enabling the possibility to exploit both the strengths and the weaknesses of the speech recognizer in deciding about the true user goal.

The performance of the model was tested in experiments with real data from the database, collected during the deployment period of the tour-guide robot RoboX at Expo.02. The results show that the Bayesian networks provide a promising probabilistic framework for error handling in multimodal dialogue systems of autonomous tour-guide robots.

9.2 Multimodal repairs in spoken human-robot interaction

9.2.1 Multimodal repair strategies using decision networks

While probabilistic modality fusion can reduce the need for repair dialogues, repair actions are still needed in the case of undefined user goal in the robot dialogue. The undefined user goals often result from adverse acoustic conditions or uncooperative user behaviors. In such conditions, the repair actions can also exploit non-speech based modalities.

In our methodological concept for designing and implementing of dialogue repair strategies, in the case of undefined user goal, the dialogue repair sequences were chosen in accordance with the tour-guide requirements, exploiting different input and output robot modalities, e.g. speech or buttons-based input, move event, etc. Given that the robot can have only a probability for the possible user goals during dialogue, the strategies for repair-action selection were modelled introducing concepts from probability and decision theories and related graphical representations, e.g. Bayesian networks and their extensions - decision networks.

The use of decision theory allowed us to define the tour-guide dialogue as a sequential process of decision-making, where decision networks were used to choose from the available actions at each dialogue state. Decision networks utilize a mathematical framework for choosing actions, based on the maximum expected utility (MEU) of the repair actions over the distribution of the user

goals given by the Bayesian network. The MEU principle allows modelling of complex task-oriented tour-guide robot behaviors, through manipulating the utility function values.

9.2.2 Multimodal repair strategies based on grounding

While the repair strategy, i.e. the sequence of repair actions, can be straightforward with two input modalities (e.g. speech and laser), incorporating new modalities would require more systematic approach in designing time-consuming repairs. For this purpose we have introduced a multimodal state-based model for grounding conversation in the general case of service robots under noisy acoustic conditions. The model exploits the multiple modalities available in the service robot system to provide evidence for reaching grounding states. The initial two states are related to the events of presence of a user who is attending to the conversation with the robot. A Bayesian network combining information from the laser and video modality was used to estimate the probabilities that the grounding states have been reached. The remaining two states in the grounding model were related to the grounding state of reliable speech modality and the grounding state of valid user goal, i.e. a user goal that can be mapped into a service provided by the robot. The speech modality reliability was explicitly modelled by the event of error in the user goal identification based on the observed recognition result. Another Bayesian network was used to model the dependencies between the event of speech modality reliability, the user goal and the speech recognition result as well as the signal-domain measure related to the level of acoustic noise.

The criterion used to consider the conversation as grounded at each particular grounding state was based on the probability of he grounding state-related events, estimated by the Bayesian networks. The use of two distinct phases of grounding has allowed us to utilize special topologies in the Bayesian networks that resulted in a reduced number of computations needed for the probabilistic inference. In particular, using a polytree (singly-connected) BN topology in the first grounding phase has allowed reduction from exponential to linear number of operations in the number of used modalities, needed by inference.

9.2.3 Evaluation of multimodal grounding in human-robot interaction

The evaluation of the error handling methods based on grounding, was performed using technical and usability evaluation. Technical evaluation was based on measures of accuracy on the component level of grounding-state prediction, as well as global measures of dialogue task success. We have defined two global measures for tour-guide dialogue success, i.e. task success and user attendance rate.

In the user-based evaluation we correlate subjective measures of performance derived from a user survey with the technical measures for dialogue task success. Although the limited amount of data, based on 22 users, the experiment outlined that the subjective measures of system usability correlate more strongly with the attendance rate of the user and less strongly with the user goal identification accuracy before and after executing any repair actions.

Both technical and the user-based evaluation supported the fact that the proposed grounding architecture can contribute to a significant gain in the accuracy of the final user goal identification, as well as a gain in user attendance rate. Hence, the use of multimodal grounding can enhance the usability of the service robot interactive system.

9.3 Future perspectives

The proposed methodology for voice interface design and error handling was done for the case of tour-guide robot. The tour-guide robot in our case provided tours in a mass exhibition. However, the solutions proposed in the thesis can be easily applied to robots that can guide people and provide information in museums, hotels, shops, airports and hospitals. A tour-guide robot can be very useful for example in aiding blind people find their way in such places. Voice-enabled interface can be also very useful in the case of edutainment and entertainment robots interacting with people without robotics knowledge in their every day environment.

In the future the work presented in the thesis can be extended in several directions:

- ♢ Given that our main objective was to clearly demonstrate that integrating speech recognition in the speech modality of a tour-guide robot operating in very noisy environment can enhance human-robot interaction, we investigated interactive system relying on restricted speech recognition solutions (e.g. system driven dialogue based on a small vocabulary of meaningful keywords). Service robots are expected to have a profound impact in the aging society of the future, and for that purpose human-robot interaction have to be studied in the context of long-term social interaction with its user. Therefore, it is straightforward to extend the dialogue scenario presented in this thesis in the direction of more natural interactive scenarios, including mixed-initiative dialogue, and incorporating recognition of more unrestricted spoken input. In enabling such interaction scenarios, investigating multimodal approaches (e.g audio-visual) to speech recognition of noisy spontaneous speech becomes very important.

- ♢ Designing more sophisticated human-robot voice-enabled interfaces will pose questions related to the minimal set of input modality sensors to be used in dialogue. In particular the methods for dialogue error handling will require additional grounding states. In this direction, existing work in dialogue strategies for resolving communication errors based on grounding can be further investigated.

- ♢ In designing new repair sequences for more complex interaction with service robots, exploiting further the use of dynamic utilities in repair action selection will be beneficial to avoid repetitive and failing repairs. In this context existing work in the field of reinforcement learning can be beneficial. Utility-based approaches can be combined with systematic repair strategies based on grounding and simple strategies like the ”curiosity” principle presented in Section 7.8.3 of the thesis. Such combined approaches could lead to better robustness of the repair strategy.

- ♢ Investigating further strategy for efficient inference with Bayesian networks would be also a requirement in more complex models of grounding. In these models decomposable networks as well as approximate inference may lead to inference computation that can be scaled efficiently with the network size. Bayesian networks can be further investigated in composing and adapting models that can automatically learn new modality events in the multimodal data such as new vocabulary words, new user identity, etc. In this case automatic structure learning and confidence measures for model composition and adaptive learning would be under focus for future research.

Finally the idea of multimodal grounding can be applied in the more general context of improving the robustness of any interactive system, including any system for interactive data acquisition from multiple modalities. For example applying utilities and grounding in multimodal authentication systems can bring better user acceptance when using spoken repairs.

Microphone Array DA-400 2.0 specifications

A

DA-400 2.0

Desktop Array™

DSDA® 2.0 DIGITAL SUPER DIRECTIONAL ARRAY

Pureaudio™

- Far-field digital microphone system incorporates new, high performance Digital Super Directional Array (DSDA®) 2.0 and PureAudio™ software to eliminate noise and enhance speech in noisy home and office environments.
- Users can interact headset-free with speech-driven desktop applications by distances of up to four feet.
- State-of-the-art software system features a DSDA adaptive beamforming technique, revolutionary de-reverberation process, and PureAudio to reduce latency and provide significantly less digital residual distortion.
- The result is a robust audio interface offering superior sensitivity and highly aggressive noise reduction for all voice-enabled desktop applications.

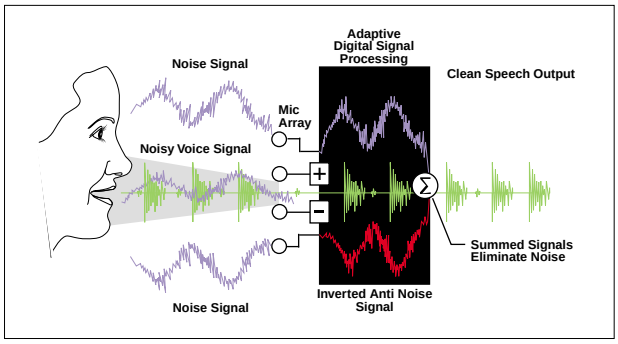
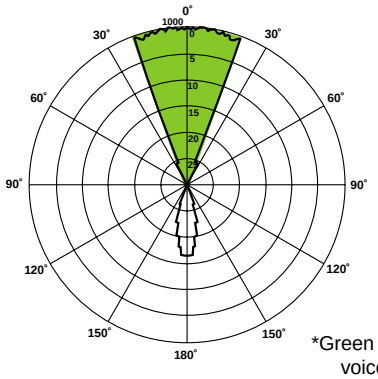


Figure 1. DSDA Illustration



*Green represents voice signal

Figure 2.
Polar Plot of Directional Microphone Sensitivity Beam
(1/3 Octave Noise, Centered at 1 kHz)

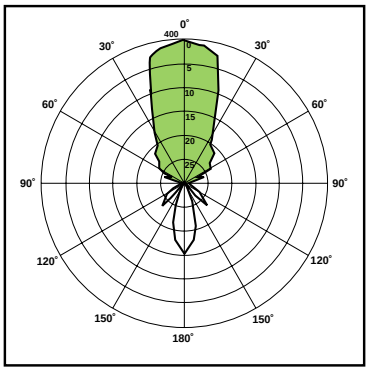
Andrea Electronics Corporation
45 Melville Park Road, Melville, New York 11747
Websites: www.AndreaElectronics.com
www.AudioCommander.com
Phone/Fax: (800) 442-7787

DSDA® 2.0
DIGITAL SUPER DIRECTIONAL ARRAY

Far-Field Microphone Technology

Digital Super Directional Array (DSDA® 2.0)

DSDA 2.0 is a sophisticated and robust noise-cancellation solution developed to bring a new level of clarity to voice communication applications. A unique feature of the technology is its ability to be embedded into speech-enabled hardware devices and enhance speech communications software products. DSDA 2.0 is adaptive and capable of being customized for a wide range of applications, so its superior noise cancellation capabilities can benefit not only users of desktop speech communications, but also users of any speech-enabled application ranging from an in-vehicle communication system to an Internet appliance to a wireless mobile communication device.



Polar Plot - DA-400 2.0 1/3 Octave @ 1 KHz

Patented DSDA adaptive microphone technology enables the optimal performance of headset-free, far-field voice input by creating a narrow reception cone of microphone sensitivity on the user's voice and canceling noise outside of that signal. DSDA version 2.0 utilizes a unique de-reverberation technique which dramatically reduces reverberation noise caused when a speaker's voice reverberates from walls or ceilings, which has the effect of degrading the performance of speech recognition applications. As a result, this software offers greater sensitivity and a superior solution for clear voice recognition with untethered, far-field voice communications.

Market Applications

Automotive: Telematics, AutoPCs, Mobile Multimedia Systems, Hands-Free Carphone Kits, Global Positioning Systems (GPS), etc.
Desktop: Speech Recognition, Internet Telephony, Videoconferencing, Voice Verification
Embedded Devices: Handheld PDA's, Set-top boxes, Professional Audio Systems, Surveillance devices, Intercoms (Home Automation), Hearing Aids, Interactive Kiosks, etc.

Specifications:

Adaptive Beamforming	2-8 Microphones
Flexible Array Structure	
Sharp Noise Reduction	Outside of a Reception Cone
Wide Tailored Frequency Range	Within 0-20 kHz
No Effect on Audio Quality	
Typical Bandwidth	0-16 kHz
Recommended Range of Operation	2' - 4'

ANDREA

Andrea Electronics Corporation
45 Melville Park Road, Melville, New York 11747
www.AndreaElectronics.com
www.AudioCommander.com
Phone/Fax: (800) 442-7787

The speech recognition system of RoboX at Expo.02

B

This appendix goes in the design details of the recognition system of RoboX. We can define the recognition system to be a short vocabulary, isolated words, speaker-independent, speech recognition system (Gibbon et al., 1997). The small set of dictionary words, implies also simpler grammar, i.e a loop of words. The recognition models is sub-word based, and introducing new words to the recognition vocabulary is trivial. The system is intended to recognize a limited vocabulary, but can accept an unlimited vocabulary input. In such a system, we are also interested in the rejection of irrelevant words. For implementation, we use the Hidden Markov Model toolkit (HTK) (Young et al., 2002). In the sections below a short introduction is given to the HMM speech recognition framework. Then we describe the experiments made with the phoneme based HMM recognition system (flexible vocabulary approach). The results are given only for the English words. The same basic steps can be applied for any other language.

Out-of-vocabulary words and spontaneous speech phenomena like breath, coughs and all other sounds that could cause a wrong interpretation of visitor's input have also to be detected and excluded. For this reason a word spotting techniques with garbage models have been added to the recognition system. At the end of the chapter the final system is tested with noisy speech files.

B.1 HMMs - basic principles

Speech is a realization of some message encoded as a sequence of one or more symbols. Recognition is the reverse operation, i. e. to recognize the underlying symbol sequence given a spoken utterance. As a first step continuous speech is converted to a sequence of equally spaced discrete parameter vectors. The sequence of vectors forms an exact representation of the speech waveform on the basis that for the duration covered by a single vector (10 ms typically), the waveform can be regarded

as being stationary. The role of the recognizer is to give a mapping between sequences of speech vectors and the wanted underlying symbol sequence. The issues that make the mapping difficult are the speech variability and the boundaries between symbols. In our case we have a limited speech input, and the expected answer is YES / NO / PYGMALION.

B.1.1 Isolated word recognition

Let each spoken word be represented by a sequence of speech vectors or observations O, defined as:

$$O = o_1, o_2, ..., o_T, \tag{B.1}$$

where o_t is the speech feature vector observed at time t. The isolated word recognition problem can then be regarded as that of computing:

$$\arg\max_i \{P(w_i|O)\}, \tag{B.2}$$

where w_i is the i'th vocabulary word. This probability is not computable directly, but using Bayes'Rule gives:

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)} \tag{B.3}$$

Thus, for a given set of prior probabilities $P(w_i)$, the most probable spoken word depends only on the likelihood $P(O|w_i)$. Given the dimensionality of the observation sequence O, the direct estimation of the joint conditional probability $P(o_1, o_2, ..., o_T|w_i)$ from examples of spoken words is not practicable. However, if a parametric model of word production such as a Markov model is assumed, then estimation from data is possible since the problem of estimating the class conditional observation densities $P(O|w_i)$ is replaced by the much simpler problem of estimating the Markov model parameters. If we look once again at Equation B.3 we see that we have also the prior probability $P(w_i)$, which can be in our case the probability of YES and NO answers. We assume that the sequence of observed speech vectors corresponding to each word is generated by a Markov model. A Markov model is a finite state machine which changes state once every time unit and each time t that a state j is entered, a speech vector o_t is generated from the emission probability density $b_j(o_t)$. Furthermore, the transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}. If we have a Markov model M and we know the state sequence and associated observation to this model it is easy to calculate $P(O, X|M)$. The joint probability that O is generated by the model M moving through the state sequence X is calculated simply as the product of the transition probabilities and the emission probabilities. However, in practice, only the observation sequence O is known and the underlying state sequence X is hidden. This is why it is called a Hidden Markov Model. Given that X is unknown, the required likelihood is computed by summing over all possible state sequences $X = x(1), x(2), x(3), ..., x(T)$, that is

$$P(O|M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t).a_{x(t)x(t+1)}, \tag{B.4}$$

where $x(0)$ is constrained to be the model entry state and $x(T+1)$ is constrained to be the model exit state. As an alternative to the equation above, the likelihood can be approximated by only considering the most likely state sequence, that is

$$\hat{P}(O|M) = \max_X \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t).a_{x(t)x(t+1)} \right\} \tag{B.5}$$

Although the direct computation of Equations B.4 and B.5 is not tractable, simple recursive procedures exist which allow both quantities to be calculated very efficiently. If Equation B.2 is computable then the recognition problem is solved.

Given a set of models M_i corresponding to words w_i, Equation B.2 is solved by using Equation B.3 and assuming that

$$P(O|w_i) = P(O|M_i) \tag{B.6}$$

All this, of course, assumes that the parameters (a_{ij}) and ($b_j(o_t)$) are known for each model M_i. Herein lies the power of the HMM framework. Given a set of training examples corresponding to a particular model, the parameters of that model can be determined automatically by a robust and efficient re-estimation procedure. Thus, provided that a sufficient number of representative examples of each word can be collected then a HMM can be constructed which implicitly models all of the many sources of variability inherent in real speech. Firstly, a HMM is trained for each vocabulary word using a number of examples of that word. Secondly, to recognize some unknown word, the likelihood of each model generating that word is calculated and the most likely model identifies the word.

B.1.2 Emission probability specification

Before the problem of parameter estimation can be discussed in more detail, the form of the emission probability distributions $\{b_j(o_t)\}$ needs to be made explicit. HTK is designed primarily for modelling continuous parameters, using continuous density multivariate distributions. It can also handle observation sequences consisting of discrete symbols in which case, the emission probability distributions are discrete probabilities. We will use the most frequent case - continuous density distributions. In HMM systems, the most common representation of the emission probability distributions are modelled with Gaussian Mixture Densities. So like parameters we then have means variances, mixture weights and transition probabilities. HTK allows each observation to be split into a S number of independent streams. The formula for computing $b_j(o_t)$ is then

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} N(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s}, \tag{B.7}$$

where M_s is the number of mixture components in stream s, c_{jsm} is the weight of the m'th component and $N(o_{st}; \mu_{jsm}, \Sigma_{jsm})$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

$$N(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)' \Sigma^{-1} (o-\mu)}, \tag{B.8}$$

where n is the dimensionality of o. The exponent γ_s is a stream weight. It can be used to give a particular stream more emphasis.

B.1.3 Algorithms for training and decoding

Our goal during training is to estimate the parameters of the HMM model and use the estimated values to do the recognition. We have two kinds of training supervised and unsupervised. The difference between them is in the presence of time-labelling information. With supervised training we have this labelling, i. e. we know the starting point and the duration of the word or the sub-word units, and initialization and training is easier. In practice, when we have a lot of training data it is more likely to use unsupervised training. In that case we can initialize the model by labelling only a short part of the train data, or by using good guess (uniform distribution of the vectors). Once

this is done, more accurate (in the maximum likelihood sense) parameters can be found by applying the so-called Baum-Welch re-estimation formulae (Young et al., 2002).

For recognition we can use Viterbi Decoding. In practice, it is preferable to base recognition on the maximum likelihood state sequence since this generalizes easily to the continuous speech case whereas the use of the total probability does not. This likelihood is computed using the following recursion (log likelihoods are used in the equation (Young et al., 1989)):

$$\psi_j(t) = \max_i\{\psi_i(t-1) + \log(a_{ij})\} + \log(b_j(o_t)). \tag{B.9}$$

This recursion forms the basis of the so-called Viterbi algorithm. This algorithm can be visualized as finding the best path through a matrix where the vertical dimension represents the states of the HMM and the horizontal dimension represents the frames of speech (i.e. time). The log probability of any path is computed simply by summing the log transition probabilities and the log output probabilities along that path. The paths are grown from left-to-right column-by-column. At time t, each partial path is known for all states i, hence the equation above can be used to compute $\psi_j(t)$ thereby extending the partial paths by one time frame. This concept of a path can be generalized to deal with the continuous speech case (Young et al., 1989).

B.2 HMM based speech recognition systems developed with HTK

In this section we provide information about the set of experiments in developing the final recognition system used for RoboX. We describe the HMM model, the training method and database Finally for each experiment we present test results and some discussion in the form of conclusion.

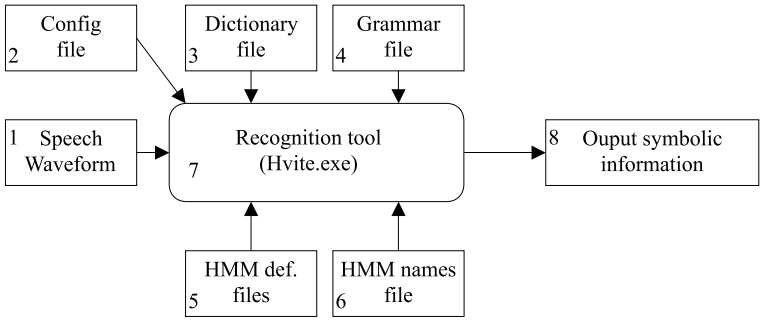
B.2.1 Speech features

It is important, when choosing the speech features, to take into account the adverse noise conditions. Even if we assume a given form of noise (for example "babble" in our case) we can have a lot of variability during the different days of the Expo.02. In that case, we should concentrate on finding noise resistant speech features - features that does not depend on noise. In general speech representation in the cepstral domain can increase the performance of recognizers in adverse noisy conditions , with respect to standard DFT (Huang et al., 2001). Choosing mel-frequency cepstral coefficients (MFCCs) as speech features is reasonable, both for robustness and because of their frequent use in recognizers.

One feature of many background noises and distortions found to occur alongside speech is that they vary slowly relative to speech. A simple method to remove the slow variations with MFCCs is to use Cepstrum Mean Normalization. This involves removing the mean of all cepstral vectors and has found to improve recognition significantly, especially that in presence of channel distortion (such as that caused by microphone changes), without degrading the baseline system. We can also remove the first (log-energy based) cepstral coefficient (C0) to reduce the mismatch between training and testing waveforms energy. In the recognition system of RoboX, we decide to use MFCCs. Future work in this direction should confirm our decision or show better ways. We can experiment with features using auditory modelling, like PLP and RASTA - PLP. Also Cepstral Mean Normalization can be performed, as a simple method for removing the slow variations, introduced by noise.

B.2.2 Description of the recognizer

The operation of the recognition tool provided with HTK can be seen from the block scheme, depicted below.




```
Command> hvite -C config2 -H hmm9/macros -H hmm9/hmmdefs -i recount.mlf -w wdnet
-p 0.0 -s 5.0 dict monophones1
```

Figure B.1: Block scheme of the recognition tool Hvite.exe, supplied with HTK

In the figure we have also an example command line used to run the recognizer. It requires for input a configuration file $config2$, macros file *macros* (it includes some common thresholds and definitions - the size of the feature vector, a floor thresholds for the model variances and etc.), HMM definition file $hmmdefs$ (in this case both *macro* and $hmmdefs$ are in the directory $hmm9$), recognition network file *wdnet*, dictionary file *dict*, file containing the names of the active HMM models $monophones1$ and the output file is $recount.mlf$ (Young et al., 2002). Bellow each block is explained and example of the files discussed are given.

Input

♢ **Speech waveform**. Can be a "live" speech directly acquired from the microphone used, or a set of test wav file.

♢ **Config file**. It is a text file in which the parameters for the initial preprocessing and functioning of the recognizer are given. For a clear idea bellow we give an example of such a file:

... # Waveform capture

SOURCERATE=625.0 # sampling period in 100ns (fs=16kHz)

SOURCEKIND=HAUDIO # source is live audio

SOURCEFORMAT=HTK # source format will be HTK

ENORMALISE=F # energy normalization (false for live audio)

USESILDET=T # use automatic silence detection (true)

MEASURESIL=F # measure background noise prior to sampling

OUTSILWARN=T # on start up message for measuring noise ...

♢ **Dictionary**. When talking about dictionary and vocabulary, in this report, we will make a small difference, although in general, it is not a common rule. Vocabulary will stand for the

set of words, which the recognizer is configured to decode. Dictionary will be used in a wider sense - it will include many words, some of them in the vocabulary some them out of it. In a phonetic dictionary we will have the set of words along with their orthographic transcription. In many cases the dictionary and the vocabulary will be the same. An example of a short dictionary file, which can be used in HTK is given bellow:

...

NO n ow sp

PYGMALION p ih g m ey l ia n sp

ROBOT r ow b oh t sp

SENT-END [ ] sil

SENT-START [ ] sil

YES y eh s sp

silence [ ] sil

...

where *sp* stands for short silence, silence is the default dictionary entry for silence *sil* is long silence. SENT-START and SENT-END are reserved words for the initial and final silence of the utterance, being recognized. This is an example dictionary, in which we have included also the word ROBOT.

♢ **Grammar**. HTK provides a grammar definition language for specifying simple task grammars such as in our task. It consists of a set of variable definitions followed by a regular expression describing the words to recognize. For our application, a suitable grammar file *gramrob* might be:

$name = PYGMALION|ROBOT;

$answer = YES|NO;

(SENT-START ($name $answer|$answer|$name) SENT-END),

where the vertical bars denote alternatives, the square brackets denote optional items and the angle braces denote one or more repetitions. The complete grammar can be depicted as a network as shown in Figure B.2.

By default, all arcs in the grammar are equally likely. However, we can attach an additional log transition probability $l = x$ to an arc. The recogniser simply adds the scaled log probability x to the path score and hence it can be regarded as an additive word transition penalty.

♢ **HMM definition files** ($hmmdefs$) . This is a set of one or more files defining HMM topology and parameters' values. At first we start with a prototype HMM file, which aim is only to define the topology. The initial values of the parameters are not important. This prototype file is used in building the later definition files. Then the appropriate training process is performed in steps (re-estimations). For each word to be recognized, a distinct HMM is used. The most common practice is to use a prototype HMM (3 state, left to right see Figure B.3), corresponding to a phoneme.

♢ **Active HMM models' name file**. This is a file containing the names of the active HMM models (used during recognition). This corresponds to the name of HMMs mapped to the words in a whole-word model, or the names of the sub-word based HMMs in the other model. The idea is that you can have a large dictionary and HMM definitions file. In the last you can

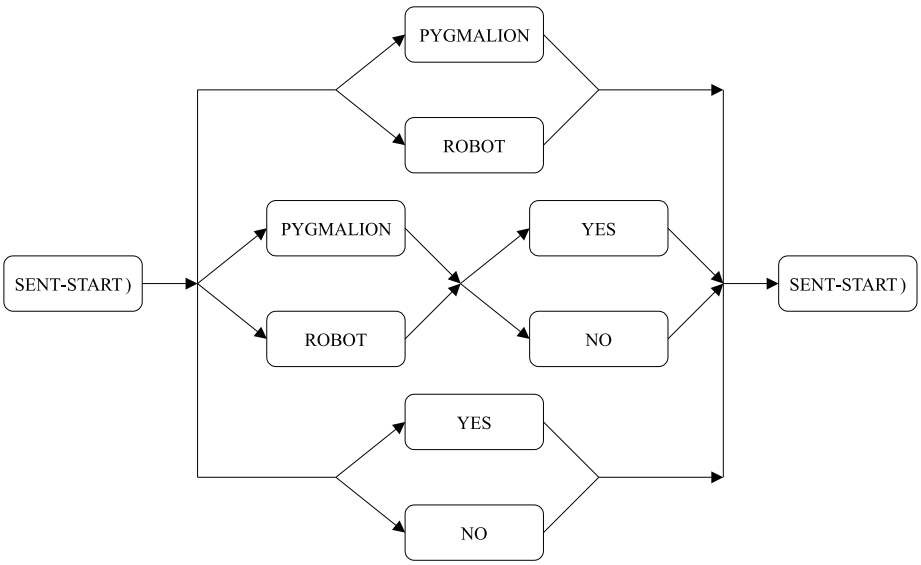


Figure B.2: Grammar for the recognizer

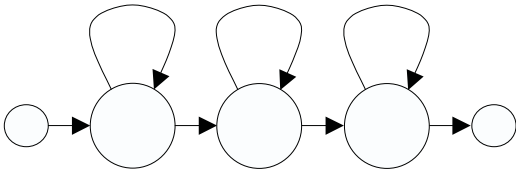


Figure B.3: Prototype HMM

even give both the word and sub-word based HMMs, but in HMMs name file you have only the names of the used in recognition models, defined by the Grammar (vocabulary). So you can play with different representations just changing the grammar and this file.

♢ 7. **Recognition tool (Hvite.exe)**. Hvite.exe is a general-purpose Viterbi word recogniser. It will match a speech file against a network of HMMs and output a transcription for each. It uses all the inputs, defined above to perform recognition on the incoming speech utterance. The output can be directly shown on the screen, when operating live, or stored in a file (recount.mlf). Depending on the configuration, it can use an automatic silence detector, and auto replay the recognized speech.

Output

The output contains symbolic information. It gives the recognized words, according to the grammar file. Additional information about average likelihood per frame and whole likelihood is supplied. If configured the recognizer gives also time duration information. All this information can be stored in a file, and outputted to the screen in ”live” mode.

B.2.3 Databases

We have recorded two databases. One was recorded with a head set microphone. The other was recorded with the microphone array. Two different rooms were used. The sampling frequency is

16kHz.

♢ **D1**. The first database is composed of 12 speakers - 7 male, 5 female. It consists of 40 prompts generated randomly from the grammar on Figure B.2. It was intended for testing. It is labelled with time duration of the words. Information about the word distribution in the set of 40 utterances is given in Figure B.4 below.

Count of name		
name	Total	prob
NO	5	0.125
PYGMALION	9	0.225
PYGMALION NO	3	0.075
PYGMALION YES	3	0.075
ROBOT	8	0.2
ROBOT NO	2	0.05
ROBOT YES	5	0.125
YES	5	0.125
Grand Total	40	1

Figure B.4: Statistics about word distribution in the Testing database (per speaker)

♢ **D2**. The second database is composed of 20 speakers (10 male 10 female) 50 prompts (Figure B.5 (b)) per speaker generated from the same grammar, but here we have only one alternative for the robot's name - Pygmalion. It was intended for training.

Count of prompts		
prompts	Total	prob
NO	48	0.24
PYGMALION	18	0.09
PYGMALION NO	15	0.075
PYGMALION YES	10	0.05
ROBOT	17	0.085
ROBOT NO	12	0.06
ROBOT YES	25	0.125
YES	55	0.275
Grand Total	200	1

(a)

Count of name		
name	Total	prob
NO	13	0.26
PYGMALION	11	0.22
PYGMALION NO	8	0.16
PYGMALION YES	8	0.16
YES	10	0.2
Grand Total	50	1

(b)

Figure B.5: Statistics about word distribution in a) D3 and b) D2 (per speaker)

♢ **D3**. In addition 200 utterances (Figure B.5 (a))of the speaker Plamen were recorded. Different parts from these databases were used in the experiments.

B.3 Description of the recognition system

The recognition system of RoboX was build using a flexible vocabulary system. In such a system instead of training a distinct HMM for each word, we train a HMM for each phoneme in the phoneme inventory of the words in the recognizer's grammar. At the end the recognized words models are composed of the corresponding sub-word models, i.e. the HMMs for each phoneme in the word. Thus, in the recognizer we train a set of phonemes, corresponding to the number of phonemes in the training utterances.

We start with single Gaussian per state and 3 state HMMs for the phonemes. We use MFCC delta and acceleration coefficients as speech features. The initial prototype HMMs are gradually trained to end up with four Gausian mixtures, three state HMMs for every phoneme. The grammar is presented in Figure B.2.

We use **D2** databse for training (10 female and 10 male speakers). For testing we use 146 utterances **D3** and 11 speakers from **D1** (7 male, 4 female) 25 utterances per speaker. We have 521 utterances in total.

```
====================== HTK Results Analysis =======================
  Date: Fri Nov 16 18:24:18 2001
  Ref : doc\testref.mlf
  Rec : recount1.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=99.52 [H=419, S=2, N=421]
WORD: %Corr=99.61, Acc=99.61 [H=510, D=1, S=1, I=0, N=512]
===================================================================
SENT is for sentences (H correct, S errors, N number of sentences)
WORD is for words (Acc % correct H correct, D deletion errors, S
substitution errors, I insertion errors, N number of words).
```

Figure B.6: Recognition performance results

B.4 Word spotting system

In the interactive system of RoboX it is not necessary, nor possible to recognize all the words pronounced by the user. First we can't predict each word, the unprepared user can utter, and second - the information we need is completely contained in our vocabulary (YES | NO | PYGMALION). Therefore, we finally use a word-spotter as a recognition system.

B.4.1 Description

Word spotting aims to detect and recognize a limited number of keywords (KWs) in the incoming speech. To detect and filter all words and speech events that are not KWs we use filter or garbage models. There are different approaches for building garbage models and for word spotting in general. The model can be a HMM based model (Wilpon et al., 1990; Renevey et al., 1997), or it can be organized without any filter models (Caminero et al., 1996; Silaghi, 2005). In our case (a few words for recognition), we can chose a method which is the most convenient for the design process, so that we introduce only minor changes in the recognition system. In that sense, HMM based garbage model is better. We also don't want additional training, if possible. We have already well trained HMM models for the phonemes. We can put them in a parallel network, which is looped. Every possible word can be modeled by a sequence of phonemes from the set of HMMs we have, even when it is not the complete language phoneme set. Such a model becomes very likely during recognition

decoding and can compete with and often outscore the keywords. One way to overcome this effect is to create more general models (for nasals, vowels and etc - (Renevey et al., 1997)) and to retrain including these models in the HMMs' set. To avoid any additional retraining we can add a penalty (a small negative number to the log-likelihood score), before our garbage model (Figure B.7). We define our filter (garbage) model as a sub-network of the recognition grammar network. This sub-network consists of all the phonemes in parallel with initial penalty as described above to weaken the initial garbage model. The value of this penalty is decided after experimentation. Pictorially our sub-network looks like the one in Figure B.7.

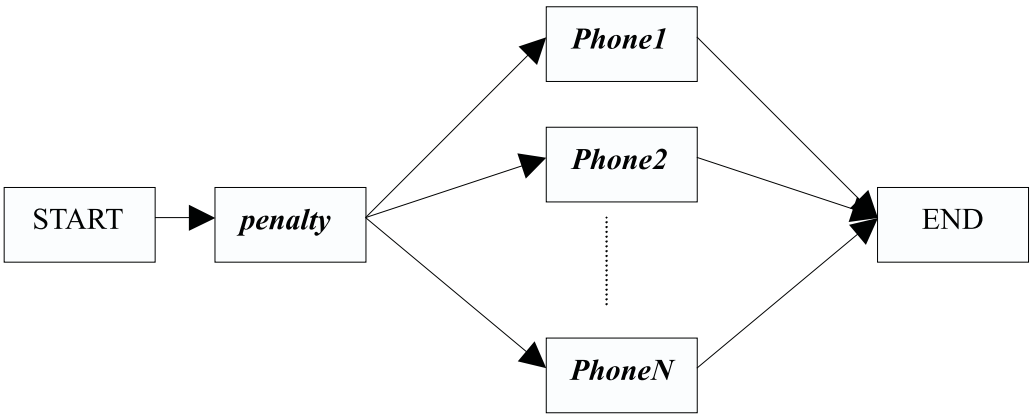


Figure B.7: Garbage recognition network (all phonemes in parallel)

B.4.2 Definitions

Definition 9 (True Hit (TH)) *: When a word in the recognized utterance is correctly spotted, we declare it as a True Hit.*

The recognizer usually outputs the word name and its start and end points in the utterance. We calculate the position of the word by summing the start and end time and dividing by two. To check the accuracy we must have a label file with the recognized utterance transcription and words' time duration. The word is correctly spotted, when it is the same as in the transcription and its position is between the start and end point of label word. In the other case we have a False Alarm.

Definition 10 (Accuracy (Acc)) *: It is the number of the Hits divided by the number of actual occurrences of the word in the test utterances.*

To evaluate a Word Spotting system we look both at the number of false alarms and accuracy for the duration of the test utterances. We should choose such a value for the penalty in our model, for which the number of false alarms is small, while keeping accuracy enough high.

Definition 11 (ROC (receiver operating characteristic)) *It is a plot of Acc (accuracy) vs. FAs (false larms) for different operating conditions to be chosen, based on the trade-off between these two parameters (in our case it is the penalty).*

Definition 12 (FOM (figure of merit)) *Average percentage of correctly spotted keywords, computed with the ROC curve between 0 and 10 FA/(KW*HR), where HR are the hours of the speech, FA are the number of false alarms, and KW are the number of keywords.*

We use directly the results about FAs and Acc with respect to penalty l in our experiments.

The word spotting system

We use the recognition system described in Section B.3. Testing is done with the prompts of 12 speakers in **D1**: 7 male (m) and 5 female (f) speakers. Duration of test data is 0.16 hours. l is the penalty added to the filtering models. When $l = 0$ garbage models are most of the time selected (very strong filtering). More negative l gives less efficiency to garbage filtering, and keywords can compete with GB words. FOM is averaged over the duration of the test data, (0.16 hours in our case). The results from the test and the chosen l value for all keywords are given in Table B.1.

- L	word	HITs	FAs	Actual	FOM%	Acc
2	Overall:	333	63	456	50.87	73%
3	Overall:	374	73	456	56.59	82%
4	Overall:	397	85	456	49	87%
5	Overall:	408	88	456	45.98	89%
6	Overall:	418	**89**	456	39.79	**92%**

Table B.1: Keyword spotting statistics (HITs - true hits, FAs - false alarms, FOM figure of merit, Acc -accuracy) for different penalties (L)

We finally choose a value of l= -6, because the FAs rate does not change significantly while the achieving good gain in accuracy.

User satisfaction tests survey results

C

Participant # _____

Personal Information:

Gender: [ ] Male [ ] Female

Age: [ ]<18 [ ] 18-24 [ ] 25-35 [ ] 36-45 [ ] 46-55 [ ] 55+

Occupation: _____

Is English your native language: [ ] Yes [ ] No

Familiarity with robots:

Have you ever used a real robot: [ ] yes [ ] no
Controlled a robot with voice: [ ] yes [ ] no
Used speech recognition software: [ ] yes [ ] no

You will now interact with the autonomous tour-guide robot RoboX. He will ask you questions and you will answer to him using single words. Please keep in mind that you will have to use single words. The moment for answering will be indicated in the eye of the robot in the form of animation (flashing ear). Please wait for this animation to start, if you want to be heard by the robot.
Otherwise, you are free to behave naturally with the robot and even experiment with him if you feel like.
Have a good time!

Figure C.1: User satisfaction survey questions, part 1 out of 3

Survey

The dialogue system in general

1. On a scale from 1 to 5 rate your satisfaction from the dialogue you had with RoboX. Cross the applicable answer:
   [ ] Boring; [ ] Not so interesting; [ ] Somehow funny; [ ] Very interested; [ ] Real fun!
2. Please, describe what did you like and did not like?
   _____
   _____
3. Did you quit or want to quit the conversation at some point? ([ ] yes; [ ] no). Why?
   _____
   _____
4. Did you experiment with the robot in some way? Please describe.
   _____
   _____
5. Did you like the style of the presentation? ([ ]yes; [ ]no).
6. Will you prefer if the robot is more serious? ([ ]yes; [ ]no).

The dialogue system components

7. On a 1 to 5 scale, rate how did RoboX understand you?
   [ ] Very bad; [ ] Bad; [ ] Satisfactory; [ ] Quite OK; [ ]Very good

RoboX is programmed to detect and react on the following situations: (1) when you are not in front of him, (2) when you don't look at him, (3) when it is very noisy, and (4) when you don't use the correct words in your answer.

8. Did you experience any of these reactions? Cross the options that apply in your case:
   [ ] RoboX reacted when I was not in front of him.
   [ ] RoboX reacted when I was not looking at him.
   [ ] RoboX reacted when it was very noisy.
   [ ] RoboX reacted when I was not using the correct words.
9. How often did RoboX used the above four reactions?
   [ ] Almoust never; [ ] Sometimes; [ ] Often; [ ] Very often
10. Was the robot correct in his reactions:
    [ ] Never; [ ] Sometimes; [ ] Often correct; [ ] Always correct

Figure C.2: User satisfaction survey questions, part 2 out of 3

11. Did these reactions help you to stay involved and more interested in the dialogue? ([ ] yes; [ ] no).
12. Did you try to provoke the robot reactions on purpose or experiment with them? ([ ] yes; [ ] no). If yes, please give details:

_____

_____

13. Which way of interacting with a robot do you prefer?
    [ ] Speech
    [ ] Buttons
    [ ] Other: specify: _____
14. Did you change your preferences during the course of dialogue?
    [ ] I liked more the use speech.
    [ ] I liked more the use of buttons.
    [ ] I liked them both.
15. Did you learn something new about the Autonomous System Lab from the presentation? Please give details.

_____

_____

16. If you were to design a dialogue scenario for RoboX what would you change?

_____

_____

17. What changes would you like to propose in order to improve the current dialogue scenario?
    [ ] No change
    [ ] Add more reactions
    [ ] Change the conversation topic. Specify:

    _____

    [ ] Other. Please give details:

    _____

    _____

18. Can you imagine other useful application for such a robot that you can recommend?

_____

_____

_____

Thank you very much for your participation.

Figure C.3: User satisfaction survey questions, part 3 out of 3

	1	2		3		4		5	6	7
No	1-5	good	bad	y/n	Reason	y/n		y/n	y/n	1-5
1	5	The rest was fun	The lecture was boring	n		y	Tried "sure" instead of "yes", normal responses, see what happens if you ignore him	y	n	2
2	4	He interacts in a very friendly way. Accurate questions		n	I was having fun	n	I did what I was told to do. I did not try to provoke him.	y	n	4
3	5	The dialogue is interesting to communicate with a machine.	The voice is slightly metalic.	n		n		y	n	5
4	3	Funny	The voice is somewhat metalic and inconprehensible	n		n		y	n	3
5	5	I like the idea to get lab info from a robot	It is a bit slow	n		n	I behaved well	y	n	4
6	3	Interaction	Simple conversation	n		n		y	n	2
7	3		I did not like "please look at me" prompt	y	To look at something else. To move faster than the robot	n		y	n	3
8	4	Interesting	Not very serious	n		n		y	n	3
9	3	I liked the conversation.	I did not like to turn around the robot when he stopped.	y	Because I wanted to test "No" respond	y	I tried to respond with full phrases	y	n	3
10	3	Funny for a short time	Boring for a second time	n		n		y	n	3
11	4	Good recognition score, I didn't have to respond twice	Limitted answers	n		n	Not really, was quite passive	y	n	4
12	4	I liked that he detected when I moved out of his sight.	I did not like that the conversation was short.	n	I did not want to quit because it was interesting for me	y	I did not look at his eyes while answering one question. He detected it and requested to look at him while answering.	y	n	4
13	4	The robot answered correctly	The sound is not so perfect	n	It is interesting to do all the experiment	n		y	n	3

Table C.1: User satisfaction survey results, table 1 of 6

No	8				9	10	11	12		13				14		
	UR	UA	NF	UG0	1-4	1-4	y/n	y/n	Comment	Sp	Btt	Otr	Comment	Sp	Btt	Both
1		1		1	3	4	y	y	Ignoring him, answeing with wrong words	1				1		
2	1	1		1	3	3	y	n		1						1
3	1	1			2	4	y	n		1	1					1
4		1		1	3	3	y	y	I tried to answer with not so standard words and it failed to understand them.		1					1
5			1		2	4	y	n		1						1
6	1	1	1		2	3	y	n		1						1
7	1	1			3	4	n	n		1				1		
8	1	1	1		2	3	n	n			1					1
9		1		1	2	4	y	n		1				1		
10		1			2	3	n	n		1				1		
11		1			2	3	y	n		1				1		
12	1	1			2	4	y	n		1				1		
13	1	1			2	4	y	n		1				1		

Table C.2: User satisfaction survey results, table 2 of 6

No	15		16	17					18	Comment
	y/n	Comment		1	2	3	4	Comment		
1	n	No, I was busy being amused by RoboX	He should speak a bit faster. Shorten the funny thing a bit.		1		1	Enable more keywords	Asking for the way. Being led ther	
2	y	After the tour I can identify the place the robot was talking about.	May be I will add some more small comentaries about the project while moving from one place to another.	1					Tourist's guide, receptionist, school guide, even for company may be?	
3	y	Yes, prof. Siegwart and his studentts helped the robot to become "alive".	0	1						
4	y	Yes,the names and locations of the secretary and the professor.	0				1	Improve the quality of the speech synthesizer and the microphone sensitivity. Adapt the microphone sensitivity to the speaker voice.		
5	y	I am familiar with the lab already	0				1	Make a dilaogue with more keywords	Blind people guide	
6	n		More interesting conversation		1				Museum or tourist guide	
7	n				1					
8	y	About the two profesors and Robota	Nothing in particular. Perhaps the question to be his friend. It is special for a lab tour.	1					For helping people to find smt in big area, like museums supermarkets etc.	
9	n	I wasn't very study	-				1	Try to have more than two possible responds	-	
10	y	I was not familiar with the wall of this lab before	-				1	Improve the speech generation, it is hardly understandable	It is fun for a short demo, but had to imagine using it in a real application as it is.	
11	y	I learned that they were involved in "guide" robots.	May be more variety in dialogue		1				Help to disabled persons	
12	y	I just learnt that the professor's office is next to the secretary's one	I would add some more possible answers		1					
13	n				1				Help blind people go somewhere in a place they don't know.	

Table C.3: User satisfaction survey results, table 3 of 6

No	1	2		3		4		5	6	7
	1-5	good	bad	y/n	Reason	y/n		y/n	y/n	1-5
14	5	I liked the way RoboX reply to me		y	To check its reactions	n		y	n	5
15	4	I liked the funny text he was saying.		n		y	I didn't look into his eyes and waited to see if he will notice it.	y	n	4
16	3	The way the robot speaks is natural and funny	Sometimes I could not understand the questions, because of poor audio quality (and my bad English ;-))	n	First time I met a speaking robot, why would I want to quit?	y	I went outside the camera area but it did not seem to influence the tour	y	n	4
17	4	I liked his funny way of asking your participation		n		n	Not really, I just went through yes or no.	y	n	4
18	4	I liked the facial gestures		n		n		y	n	4
19	3		Interaction too slow, very imperfect speech recognition	y	I knew the dialogue already	y	I kicked it	y	n	1
20	5	I liked the behaviour of the robot and its capability to navigate autonomously. Also the speech recognition part is very good.		n		n		y	n	4
21	2	I was curious to see how is it to communicate with a robot	I had to repeat too many times in order for the robot to understand	y	Because I was not sure that he was working OK	y	Changing position, trying to speak louder, answer always with no.	y	n	2
22	4	Its sense of humor	The way it turns is not intuitive	n	The experiment is interesting and funny		I had prior experience during Expo.02	y	n	3
Cnts: AVR	**3.82**			**5 23%**				**22 100%**	**0 0%**	**3.4**

Table C.4: User satisfaction survey results, table 4 of 6

No	8				9	10	11	12		13				14		
	UR	UA	NF	UG0	1-4	1-4	y/n	y/n	Comment	Sp	Btt	Otr	Comment	Sp	Btt	Both
14	1	1			2	4	y	y	I said I did not want to be his friend. His reaction was funny to me.	1						1
15		1			2	4	y	y	Not looking into the eyes	1				1		
16	1				1	3	y	y	See question 4	1				1		
17	1				2	4	y	n		1				1		
18	1	1		1	2	4	y	n		1		1	My hand and face gestures	1		
19	1				2	2	y	n		1					1	
20	1				2	4	y	n		1				1		
21	1	1	1		4	2	y	n			1				1	
22	1	1		1	3	3	y	n		1				1		
Cnts: AVR	**15 68%**	**17 77%**	**4 18%**	**6 27%**	**2.3**	**3.5**	**19 86%**	**4 18%**		**19 86%**	**4 18%**	**1 5%**		**13 59%**	**2 9%**	**7 32%**

Table C.5: User satisfaction survey results, table 5 of 6

No	15		16	17					18	Comment
	y/n	Comment		1	2	3	4	Comment		
14	n	I have done this dialogue before	He never asks something about you! Some dialogue to "make knowledge" would be nice.		1	1		Asking personal information	No	
15	y	I learnt about the team and the first experiment of RoboX	Nothing. Everything is already great.	1					Organized tour.	
16	y	I learned a bit about the robot history and also and also the fact that professor Siegwart was the "project's father".	May be ask some questions about the visitor (without necessarilly using the answer).		1				Information desk	
17	y	Working principles of RoboX		1					Museum guide, ticket controller on a train	
18	y	I got information about the other robot and its features, and the names of the professors in the project.	Friendly dialogues are nice but they are rare so there may be more. During demo these friendly dialogues were irrelevant.		1		1	It would be nice if it can react to more words. For example it could ask the name of the person and call him by his name, and it could ask some more questions and use the answers during dialogue.	The guide, receptionist.	
19	n		Better speech recognition engine, faster navigation.	1				3 times failure with speech should permanently swich to buttons permanently with current user.		
20	y	I learned about the research done in the lab	Nothing				1	Perhaps, face recognition		
21	y	The name of professor Siegwart and the location of his office. Because I was concentrated in answering properly I did not listen very carefully.	The topic was OK but the interaction has to be faster.		1		1	The new reactions should be faster.	Blind people aid	In this case the camera of RoboX was not working
22	y	I am familiar with the lab from before	The motion of the robot and position of the robot after moving should be more intutitive for the user	1				I would like to introduce a state where RoboX is asking a user to be in front of him	All types of tour-guiding	
Cnts: AVR	**15 68%**			**7 32%**	**10 45%**	**1 5%**	**8 36%**			

Table C.6: User satisfaction survey results, table 6 of 6

Bibliography

Aji, S. M., McEliece, R. J., 2000. The generalized distributive law. IEEE Transactions on Information Theory 46 (2), 325–343.

Alami, R., Oct. 2002. Diligent: Towards a personal robot. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Antoniol, G., Cattoni, R., Cettolo, B., Federico, M., 1993. Robust speech understanding for robot telecontrol.
URL citeseer.ist.psu.edu/antoniol93robust.html

Aoyama, K., Shimomura, H., April 2005. Real world speech interaction with a humanoid robot on a layered robot behavior architecture. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, ICRA05. Barcelona, Spain, pp. 3825–3830.

Arndt, O., Oct. 2002. Interactive exhibition design: Robots in exhibitions. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Balaguer, C., Casalas, A., Chatila, R., Christaller, T., Dai, I., Fiorini, P., Haegele, M., Hirzinger, G., Knoll, A., Laugier, C., Melchiorri, C., Molfino, R., Ollero, A., Prassler, E., Santos-Victor, J., Valero, P., Schilling, K., Siegwart, R., Zhang, J., 2004. Robotics research roadmap. EURON Report KA1, EURON: the European Robotics Research network.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., Behnke, S., 2005. Multimodal conversation between a humanoid robot and multiple persons. In: Proceedings of the Workshop on Modular Construction of Humanlike Intelligence at the Twentieth National Conferences on Artificial Intelligence (AAAI), Pittsburgh / USA.

Berouti, M., Schwartz, R., Makhoul, J., April 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. pp. 208–211.

Bessière, P., Ahuactzin, J.-M., Aycard, O., Bellot, D., Colas, F., Coué, C., Diard, J., Garcia, R., Koike, C., Lebeltel, O., LeHy, R., Malrait, O., Mazer, E., Mekhnacha, K., Pradalier, C., Spalanzani, A., 2003. Survey: Probabilistic methodology and techniques for artefact conception and development. Technical Report RR-4730, INRIA Rhône-Alpes, Montbonnot, France.

Bilmes, J., Kirchhoff, K., October 2000. Directed graphical models of classifier combination: Application to phone recognition. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing. Beijing.

Bischoff, R., October 8–9 1999. Natural communication and interaction with humanoid robots. In: Proceedings of the Second International Symposium on Humanoid Robots, Tokyo. pp. 121–128.
URL citeseer.ist.psu.edu/bischoff99natural.html

Bohus, D., 2004. Error awareness and recovery in task-oriented spoken dialog systems. Ph.d. thesis proposal, Computer Science Departmet, Carnegie Mellon University.

Boros, M., Eckert, W., Gallwitz, F., Görz, G., Hanrieder, G., Niemann, H., 1996. Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In: Proc. ICSLP '96. Vol. 2. Philadelphia, PA, pp. 1009–1012.
URL citeseer.ist.psu.edu/boros96towards.html

Bourgard, W., Thrahanias, P., Hahnel, D., Moors, M., Schulz, D., Baltzakis, H., Argyros, A., Oct. 2002. Tourbot and webfair: Web-operated mobile robots for tele-presence in populated exhibitions. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Brennan, S. E., Hulteen, E. A., April-June 1995. Interaction and feedback in a spoken language system: a theoretical framework. Knowledge-Based Systems 8 (2-3), 143–151.

Brito, A. C. D., Gámez, J. C., Sosa, D. H., Santana, M. C., Navarro, J. L., González, J. I., Artal, C. G., Pérez, I. P., Martel, A. F., Tejera, M. H., Rodríguez, J. M., 2001. Eldi: An agent based museum robot. Systems Science, Special Issue: Advances in Robotics: Virtual Reality, Robot Manipulators, Bipeds and Mobile Robots 27 (4).

Bulyko, I., Kirchhoff, K., Ostendorf, M., Goldberg, J., March 2005. Error-correction detection and response generation in a spoken dialogue system. Speech Communication 45 (3), 271–288.

Burgard, W., Cremers, A. B., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W., Thrun, S., 1999. Experiences with an interactive museum tour-guide robot. Artificial Intelligence 114 (1-2), 1–53.

Caminero, J., de la Torre, C., Villarrubia, L., Martín, C., Hernández, L., 1996. On-line garbage modeling with discriminant analysis for utterance verification. In: Proc. ICSLP '96. Vol. 4. Philadelphia, PA, pp. 2111–2114.
URL citeseer.ist.psu.edu/caminero96line.html

Carpenter, P., Jin, C., Wilson, D., Zhang, R., Bohus, D., Rudnicky, A., 2001. Is this conversation on track? In: Proc. of Eurospeech 2001. Aalborg, Denmark.

Cheng, J., Greiner, R., 1999. Comparing bayesian network classifiers. In: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99). Morgan Kaufmann Publishers, San Francisco, CA, pp. 101–108.

Choi, C., Kong, D., Kim, J., Bang, S., October 2003. Speech enhancement and recognition using circular microphone array for service robots. In: IEEE Int. Conf. on Robotics and Automation (ICRA'03). Las Vegas, Nevada, USA.

Clark, H. H., Schaefer, E. F., 1989. Contributing to discourse. Cognitive Science 13 (2), 259–294.

Clodic, A., Fleury, S., Alami, R., Herrb, M., Chatila, R., July 18-20 2005. Supervision and interaction: Analysis from an autonomous tour-guide robot deployment. In: proceedings of the International Conference on Advanced Robotic (ICAR). Seatle, WA, USA.

Cooper, G. F., 1990. The computational complexity of probabilistic inference using Bayesian belief networks (research note). Artif. Intell. 42 (2-3), 393–405.

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. John Wiley & sons.

Cox, S., Dasmahapatra, S., October 2002. High-level approaches to confidence estimation in speech recognition. IEEE Transactions on Speech and Audio Processing 10 (7), 460–471.

Davis, G., 2002. Noise Reduction in Speech Applications. CRC Press, Inc., Boca Raton, FL, USA.

de Mori, R., 1998. Spoken dialogues with computers. Academic Press.

Dempster, A., N.M.Laird, D.B.Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal Royal Stat. Soc., Series B 39 (1), 1–38.

Deng, L., Huang, X., 2004. Challenges in adopting speech recognition. Commun. ACM 47 (1), 69–75.

Diard, J., Bessière, P., Mazer, E., December 2003. A survey of probabilistic models, using the bayesian programming methodology as a unifying framework. In: Proc. of the Int. Conf. on Computational Intelligence, Robotics and Autonomous Systems. Singapore (SG).
URL http://emotion.inrialpes.fr/bibemotion/2003/DBM03

Drygajlo, A., Prodanov, P., Ramel, G., Messier, M., Siegwart, R., 2003. On developing voice enabled interface for interactive tour-guide robots. Advanced Robotics 17 (7), 599–616.

Dutoit, T., 1997. An introduction to text-to-speech synthesis. Kluwer Academic Publishers, Norwell, MA, USA.

Dybkjaer, L., Bernsen, N. O., Minker, W., June 2004. Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43 (1-2), 33–54.

El-Maliki, M., Drygajlo, A., September 1999. Missing features detection and estimation for robust speaker verification. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH99). Vol. 2. pp. 975–978.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech, and Signal Processing 34 (1), 52–59.

Gales, M., 1995. Model-based techniques for noise robust speech recognition. Ph.d. thesis, University of Cambridge.

Garcia-Mateo, C., Reichl, W., Ortmanns, S., December 12-15 1999. On combining confidence measures in HMM-based speech recognizers. In: Int. Workshop on Automatic Speech Recognition and Understanding, ASRU'99. Keystone, Colorado, USA.

Gibbon, D., Mertins, I., R. Moore, e., 2000. HANDBOOK OF MULTIMODAL AND SPOKEN DIALOGUE SYSTEMS: RESOURCES, TERMINOLOGY AND PRODUCT EVALUATION. Kluwer Academic Publishers.

Gibbon, D., Moore, R., R. Winski, e., 1997. Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter: Berlin, New York.

Gieselmann, P., Waibel, A., 2005. What makes human-robot dialogues struggle? In: Proc. of the Ninth Workshop on the Semantics and Pragmatics of Dialogue (DIALOR).

Graf, B., Hans, M., Schraft, R. D., 2004. Care-o-bot ii - development of a next generation robotic home assistant. Auton. Robots 16 (2), 193–205.

Graph, B., Barth, O., Oct. 2002. Entertainment robotics: Examples, key technologies and perspctives. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Haasch, A., Hohenner, S., Hüwel, S., Kleinehagenbrock, M., Lang, S., Toptsis, I., Fink, G. A., Fritsch, J., Wrede, B., Sagerer, G., May 2004. BIRON – The Bielefeld Robot Companion. In: Prassler, E., Lawitzky, G., Fiorini, P., Hägele, M. (Eds.), Proc. Int. Workshop on Advances in Service Robotics. Fraunhofer IRB Verlag, Stuttgart, Germany, pp. 27–32.

Hanebeck, U. D., Fischer, C., Schmidt, G., 1997. Roman: A mobile robotic assistant for indoor service applications. In: Proceedings of the 1997 IEEE/RSJ/GI International Conference on Intelligent Robots and Systems (IROS'97). Grenoble, Frankreich, pp. 518–525.

Hanson, B., Applebaum, T. H., 1990. Features for noise-robust speaker-independent word recognition. In: Proc. ICSLP, volume 2. pp. 550–553.

Hara, I., Asano, F., Asoh, H., Ogata, J., Ichimura, N., Kawai, Y., Kanehiro, F., Hirukawa, H., Yamamoto, K., September 28- October 02 2004. Robust speech interface based on audio and video information fusion for humanoid hrp-2. In: Proc. of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2004. pp. 2404–2410.

Hermansky, H., 1990. Perceptual linear predictive (plp) analysis of speech. Journal of the Acoustic Society of America 87 (4), 1738–1752.

Hermansky, H., Morgan, N., 1994. Rasta processing of speech. IEEE Trans. Speech and Audio Processing 2 (4), 578–589.

Hirsch, H., Meyer, P., Ruehl, H., 1991. Improved speech recognition using high-pass filtering of subband envelopes. In: Eurospeech. pp. 413–416.

Hirschberg, J., Litman, D., Swerts, M., June 2004. Prosodic and other cues to speech recognition failures. Speech Communication 43 (1-2), 155–175.

Holzapfel, H., Gieselmann, P., November 2004. A way out of dead end situations in dialogue systems for human-robot interaction. In: Proc. of the IEEE/RAS International Conference on Humanoid Robots, 2004. pp. 184 – 195.

Hong, J.-H., Song, Y.-S., Cho, S.-B., April 18-22 2005. A hierarchical bayesian network for mixed-initiative human-robot interaction. In: 2005 IEEE International Conference on Robotics and Automation, ICRA 2005. Barcelona, Spain, pp. 3819–3824.

Horswill, I., 1992. Visual support for navigation in the polly system. In: Proc. of the 1992 AAAI Fall Symposium on Applications of Artificial intelligence to Real-World Autonomous Mobile Robots. pp. 62–67.

Horvitz, E., Paek, T., 1999. A computational architecture for conversation. In: UM '99: Proceedings of the seventh international conference on User modeling. Springer-Verlag New York, Inc., Secaucus, NJ, USA, pp. 201–210.

Horvitz, E., Paek, T., November 2000. Deeplistener: Harnessing expected utility to guide clarification dialog in spoken language systems. In: ICSLP 2000: 6th International Conference on Spoken Language Processing. Beijing, CHINA.

Horvitz, E., Paek, T., 2001. Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In: UM '01: Proceedings of the 8th International Conference on User Modeling 2001. Springer-Verlag, London, UK, pp. 3–13.

Huang, X., Acero, A., Hon, H.-W., 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development, 1st edition. Prentice Hall PTR.

Huttenrauch, H., Green, A., Norman, M., Oestreicher, L., Eklundh, K., May 2004. Involving users in the design of a mobile office robot. IEEE Transactions on Systems, Man and Cybernetics, Part C 34 (2), 113–124.

Jensen, B., Froidevaux, G., Greppin, X., Lorotte, A., Mayor, L., Meisser, M., Ramel, G., Siegwart, R., Sept. - Oct. 2002a. The interactive autonomous mobile system robox. In: Int. Conf. on Intelligent Robots and Systems, IROS 2002. Lausanne, Switzerland, pp. 1221–1227.

Jensen, B., Froidevaux, G., Greppin, X., Lorotte, A., Mayor, L., Meisser, M., Ramel, G., Siegwart, R., Oct. 2002b. Visitor flow management using human-robot interaction at expo.02. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Jensen, B., Tomatis, N., Mayor, L., Drygajlo, A., Siegwart, R., December 2005. Robots meet humans - interaction in public spaces 52 (6), 1530– 1546.

Jensen, F., 1996. An Introduction to Bayesian Networks, 1st edition. UCL Press.

Jiang, H., April 2005. Confidence measures for speech recognition: A survey. Speech Communication 45 (4), 455–470.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., Saul, L. K., Nov 1999. An introduction to variational methods for graphical models. Machine Learning 37 (2), 183 – 233.

Josifovski, L., August 2002. Robust automatic speech recognition with missing and unreliable data. Ph.d. thesis, Department of Computer Science, University of Sheffield, UK.

Kam, M., Zhu, X., Kalata, P., 1997. Sensor fusion for mobile robot navigation. In: Proceedings of the IEEE 85 (1). pp. 108–119.

Kamm, C., 1994. User Interfaces for voice applications, Voice communication between humans and machines. National Academy Press, Washington, DC.

Keizer, S., den Akker, R. O., Hijholt, A., 2002. Dialogue act recognition with bayesian networks for dutch dialogues. In: Proc. of 3rd SIGdial Workshop on Discourse and Dialogue. Philadelphia, PA, USA.

Keller, E., Werner, S., June 1997. Automatic intonation extraction and generation for french. In: 14th CALICO Annual Symposium. West Point. NY.

Kittler, J., 2000. A framework for classifier fusion: Is it still needed?. In: SSPR/SPR. pp. 45–56.

Kittler, J., Matas, J., Jonsson, K., Ramos Sánchez, M., 1997. Combining evidence in personal identity verification systems. Pattern Recognition Letters 18, 845–852.

Kjaerulff, U., 1992. Optimal decomposition of probabilistic networks by simulated annealing. Statistics and Computing 2, 7–17.
URL citeseer.csail.mit.edu/rul92optimal.html

Kleinehagenbrock, M., Lang, S., Fritsch, J., Lömker, F., Fink, G. A., Sagerer, G., September 2002. Person tracking with a mobile robot based on multi-modal anchoring. In: Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN). IEEE, Berlin, Germany, pp. 423–429.

Krahmer, E., Swerts, M., Theune, M., Weegels, M., 2001. Error detection in spoken human-machine interaction. International journal of speech technology 4 (1), 19–30.

Kulyukin, V., Gharpure, C., de Graw, N., 2004. Human-computer interaction in a robotic guide for visually impaired. In: Proc. AAAI Spring Symposium. Palo Alto, CA.

Lang, S., Kleinehagenbrock, M., Hohenner, S., Fritsch, J., Fink, G. A., Sagerer, G., 2003. Providing the basis for human-robot-interaction: a multi-modal attention system for a mobile robot. In: ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces. ACM Press, New York, NY, USA, pp. 28–35.

Lauritzen, S. L., 1995. The em algorithm for graphical association models with missing data. Comput. Stat. Data Anal. 19 (2), 191–201.

Lerner, U., Parr, R., 2001. Inference in hybrid networks: Theoretical limits and practical algorithms. In: 17th Conference on Uncertainty in Artificial Intelligence. pp. 310–318.
URL citeseer.ist.psu.edu/lerner01inference.html

Li, S., Haasch, A., Wrede, B., Fritsch, J., Sagerer, G., 2005. Human-style interaction with a robot for cooperative learning of scene objects. In: ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces. ACM Press, New York, NY, USA, pp. 151–158.

Lienhart, R., Maydt, J., 2002. An extended set of haar-like features for rapid objection detection. In: IEEE ICIP. pp. 900–903.

Luperfoy, S., Duff, D., 1996. A four-step dialogue recovery program. In: Proc. of the AAAI workshop on Detection, Repair, and Prevention of Human-Machine Miscommunication.

Maeyama, S., Yuta, S., Harada, A., Oct. 2002. Mobile robots in art museum for remote appreciation via internet. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Mandic, D. P., Obradovic, D., Kuh, A., Adali, T., Trutschel, U., Golz, M., Wilde, P. D., Barria, J. A., Constantinides, A., Chambers, J. A., 2005. Data fusion for modern engineering applications: An overview. In: ICANN (2). pp. 715–721.

Matia, F., Rodriguez-Losada, D., Galan, R., Jimenez, A., Oct. 2002. Experiments at a trade fairs with blacky the robot. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Matsui, T., Asoh, H., Fry, J., Motomura, Y., Asano, F., Kurita, T., Hara, I., Otsu, N., 1999. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services. In: AAAI/IAAI. pp. 621–627.
URL citeseer.ist.psu.edu/matsui99integrated.html

McTear, M., 2004. Spoken Dialogue Technology: Towards the Conversational User Interface. Springer.

McTear, M. F., 2002. Spoken dialogue technology: enabling the conversational user interface. ACM Comput. Surv. 34 (1), 90–169.

Murphy, K., July 2002. Dynamic bayesian networks: representation, inference and learning. Ph.d. thesis, U.C. Berkeley.

Nefian, A. V., Liang, L., Pi, X., Liu, X., Murphy, K., 2002. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP Journal on Applied Signal Processing 11, 1–15.

Nilson, N., 1984. Shakey the robot. Technical Report 323, SRI International, Menlo Park, California.

Nourbakhsh, I. R., Oct. 2002. The mobot museum robot installations: A five year experiment. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Oviatt, S., 1999. Mutual disambiguation of recognition errors in a multimodel architecture. In: CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press, New York, NY, USA, pp. 576–583.

Oviatt, S., Coulston, R., Lunsford, R., 2004. When do we interact multimodally?: cognitive load and multimodal communication patterns. In: ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces. ACM Press, New York, NY, USA, pp. 129–136.

Paek, T., Horvitz, E., 1999. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In: Brennan, S. E., Giboin, A., Traum, D. (Eds.), Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems. American Association for Artificial Intelligence, Menlo Park, California, pp. 85–92.
URL citeseer.ist.psu.edu/paek99uncertainty.html

Paek, T., Horvitz, E., August 2003. On the utility of decision-theoretic hidden subdialog. In: Proceedings of International Speech Communication Association (ISCA) Workshop on Error Handling in Spoken Dialogue Systems. Chateaux d'Oex, Switzerland, pp. 95–100.

Paek, T., Horvitz, E., Ringger, E., November 2000. Continuous listening for unconstrained spoken dialog. In: ICSLP 2000: 6th International Conference on Spoken Language Processing. Beijing, CHINA.

Paskin, A., September 2004. Exploiting locality in probabilistic inference. Ph.d. thesis, University of California, Berkley.

Pavlovic, V. I., 1999. Dynamic Bayesian networks for information fusion with application to human-computer interfaces. Ph.d. thesis, University of Illinois at Urbana-Champaign.

Pernkopf, F., Bilmes, J., 2005. Discriminative versus generative parameter and structure learning of bayesian network classifiers. In: ICML '05: Proceedings of the 22nd international conference on Machine learning. ACM Press, New York, NY, USA, pp. 657–664.

Prodanov, P., Drygajlo, A., September 2003. Bayesian networks for spoken dialogue management in multimodal systems of tour-guide robots. In: Proceedings of the 8th European Conference on Speech Communication and Technology, Eurospeech 2003. Geneva, Switzerland, pp. 1057–1060.

Prodanov, P., Drygajlo, A., March 2005. Bayesian networks based multi-modality fusion for error handling in human-robot dialogues under noisy conditions. Speech Communication 45 (3), 231–248.

Prodanov, P., Drygajlo, A., Ramel, G., Messier, M., Siegwart, R., Sept. - Oct. 2002. Voice enabled interface for interactive tour-guide robots. In: Proc. Int. Conf. on Intelligent Robots and Systems, IROS 2002. Lausanne, Switzerland, pp. 1332–1337.

Raj, B., Stern, R., September 2005. Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine 22 (5), 101–116.

Rajkishore Prasad, Hiroshi Saruwatari, K. S., 2004. Robots that can hear, understand and talk. Advanced Robotics 18 (5).

Renevey, P., 2000. Speech recognition in noisy conditions using missing feature approach. Ph.D. thesis 2303, Swiss Federal institute of Technology.

Renevey, P., Drygajlo, A., April 1997. Securized flexible vocabulary voice messaging system on unix workstation with isdn connection. In: Proc. of the 5th European Conference on Speech Communication and Technology (EUROSPEECH97). pp. 1615–1619.

Renevey, P., Drygajlo, A., 2000. Introduction of a reliability measure in missing data approach for robust speech recognition. In: Proceedings of 10th European Signal Processing Conference (EUSIPCO 2000). pp. pp. 473–476.

Renevey, P., Drygajlo, A., Bornet, O., Bourlard, H., de Weck, L., 1997. Automatic french speech recognition on unix workstations with swissnet connection. Technical Report CTI Project 2998.1, LTS-EPFL, IDIAP, ACOMM.

Roy, N., Pineau, J., Thrun, S., 2000. Spoken dialogue management using probabilistic reasoning. In: Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000).

Russell, S., Norvig, P., 2003. Artificial intelligence: a modern approach, 2nd Edition. Prentice Hall.

San-Segundo, R., Pellom, B., Ward, W., Pardo, J., June 2000. Confidence measures for dialogue management in the cu communicator system. In: Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'00. pp. 1237–1240.

Schulte, J., Rosenberg, C. R., Thrun, S., 1999. Spontaneous, short-term interaction with mobile robots. In: ICRA. pp. 658–663.

Shachter, R., 1998. Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In: UAI.

Sidner, C. L., Kidd, C., Lee, C., Lesh, N., January 2004. Where to look: A study of human-robot engagement. In: Proc. Intelligent User Interfaces (IUI). Funchal, Island of Madeira, Portugal, pp. 78–84.

Siegwart, R., Arras, K. O., Bouabdallah, S., Burnier, D., Froidevaux, G., Greppin, X., Jensen, B., Lorotte, A., Mayor, L., Meisser, M., Philippsen, R., Piguet, R., Ramel, G., Terrien, G., Tomatis, N., 2003. Robox at expo.02: A large-scale installation of personal robots. Robotics and Autonomous Systems 42 (3-4), 203–222.

Silaghi, M. C., 2005. Iterative segmentation in keyword spotting: relation to fractional programming. Technical Report CS-2005-13, Florida Institute of Technology.

Silaghi, M.-C., Bourlard, H., 1999. Iterative posterior-based keyword spotting without filler models. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU'99) Workshop.

Skantze, G., August 28-31 2003. Exploring human error handling strategies: Implications for spoken dialogue systems. In: ITR Workshop on Error Handling in Spoken Dialogue Systems. Chateau d'Oex, Vaud, Switzerland, pp. 71–76.

Smith, M., 2003. Bayesian sensor fusion: A framework for using multi-modal sensors to estimate target locations and identities in a battlefield scene, ph.d. thesis. Ph.d. thesis, Department of Statistics, Florida State University.

Smith, M., Srivastava, A., 20-22 October 2004. A bayesian framework for statistical, multi-modal sensor fusion. In: Proceedings of the Tenth U.S. Army Conference on Applied Statistics.

Spiliotopoulos, D., Androutsopoulos, I., Spyropoulos, C. D., 2001. Human-robot interaction based on spoken natural language dialogue. In: Proc. European Workshop on Service and Humanoid Robots.
URL citeseer.ist.psu.edu/spiliotopoulos01humanrobot.html

Stern, R., Raj, B., Moreno, P., April 1997. Compensation for environmental degradation in automatic speech recognition. In: ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Chanels.

Sturm, J., Boves, L., March 2005. Effective error recovery strategies for multimodal form-filling applications. Speech Communication 45 (3), 289–303.

Sturm, J., Wang, F., Cranen, B., September 1-2 2001. Adding extra input/output modalities to a spoken dialogue system. In: Proc. 2nd ACL SIGdial Workshop on Discourse and Dialogue. Aalborg, Denmark, pp. 162–165.

Suhm, B., Myers, B., Waibel, A., 2001. Multimodal error correction for speech user interfaces. ACM Trans. Comput.-Hum. Interact. 8 (1), 60–98.

Thorpe, J., McEliece, R., January-March 2002. Data fusion algorithms for collaborative robotic exploration. Progress Reports 24-149, IPN.

Thrun, S., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Hahnel, D., Rosenberg, C., Roy, N., Schulte, J., Schulz, D., May 1999a. Minerva: A second generation museum tour-guide robot. In: IEEE Int. Conf. on Robotics and Automation (ICRA'99). Detroit, Michigan.

Thrun, S., Bennewitz, M., Burgard, W., Cremers, A. B., Dellaert, F., Fox, D., Hahnel, D., Lakemeyer, G., Rosenberg, C., Roy, N., Schulte, J., Schulz, D., August 1999b. Experiences with two deployed interactive tour-guide robots. In: Proceedings of the International Conference on Field and Service Robotics (FSR'99). Pittsburgh, PA.

Topp, E. A., Kragic, D., Jensfeld, P., Christensen, H. I., April 2004. An interactive interface for service robots. In: Proceedings of the 2004 IEEE International Conference on Robotics and Automation, ICRA04. New Orleans, LA, USA, pp. 3469–3474.

Torrance, M., 1994. Natural communication with mobile robots. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.
URL citeseer.ist.psu.edu/torrance94natural.html

Torres, F., Hurtado, L., García, F., Sanchis, E., Segarra, E., March 2005. Error handling in a stochastic dialog system through confidence measures. Speech Communication 45 (3), 211–229.

Traum, D., 1999. Computational models of grounding in collaborative systems. In: AAAI Fall Symposium on Psychological Models of Communication. pp. 124–131.

Traum, D. R., Dillenbourg, P., Sep. 23 1998. Towards a normative model of grounding in collaboration.
URL http://citeseer.ist.psu.edu/133698.html

Turunen, M., 2004. Jaspis - a spoken diallogue architecture and its applications. Ph.d. thesis, University of Tampere.

Turunen, M., Hakulinen, J., 2001. Agent-based error handling in spoken dialogue systems. In: Proc. Eurospeech. pp. 2189–2192.

van den Bosch, A., Krahmer, E., Swerts, M., 2001. Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches. In: Meeting of the Association for Computational Linguistics. pp. 499–506.
URL citeseer.ist.psu.edu/vandenbosch01detecting.html

Vaseghi, S. V., Milner, B. P., January 1997. Noise compensation methods for hidden markov model speech recognition in adverse environments. IEEE. Trans. on Speech and Audio Processing 5 (1), 11–21.

Vestli, S. J., Oct. 2002. Design of and operational experiences from five museum robot istallations. In: Workshop: Robotics in Exhibitions, IROS 2002. Lausanne, Switzerland.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), ISSN: 1063-6919. Vol. 1. pp. 511–518.

Walker, M. A., Wright, J. H., Langkilde, I., 2000. Using natural language processing and discourse features to identify understanding errors. In: ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1111–1118.

Wang, Y.-Y., Deng, L., Acero, A., September 2005. Spoken language understanding. IEEE Signal Processing Magazine 22 (5), 16–31.

Willeke, T., Kunz, C., Nourbakhsh, I., May 2001. The history of the mobot museum robot series: an evolutionary study. In: FLAIRS.

Wilpon, J. G., Rabiner, L. R., Lee, C.-H., Goldman, E. R., 1990. Automatic recognition of keywords in unconstrained speech using hidden markov models. IEEE Trans. On ASSP 38 (11), 1870–1878.

Woodland, P., Gales, M., Pye, D., 1996a. Improving environmental robustness in large vocabulary speech recognition. In: Proc. ICASSP '96. Atlanta, GA, pp. 65–68.
URL citeseer.ist.psu.edu/woodland96improving.html

Woodland, P., Pye, D., Gales, M., 1996b. Iterative unsupervised adaptation using maximum likelihood linear regression. In: Proc. ICSLP '96. Vol. 2. Philadelphia, PA, pp. 1133–1136.
URL citeseer.csail.mit.edu/woodland96iterative.html

Yamamoto, S., Valin, J.-M., Nakadai, K., Rouat, J., Michaud, F., Ogata, T., Okuno, H. G., April 18-22 2005. Enhanced robot speech recognition based on microphone array source separation and missing feature theory. In: Proc. of 2005 IEEE International Conference on Robotics and Automation, ICRA 2005. Barcelona, Spain, pp. 3819–3824.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. The HTK Book (Version 3.2.1).

Young, S., Russell, N., Thornton, J., July 1989. Token passing: a simple conceptual model for connected speech recognition systems. Technical Report TR.38, Cambridge University Engineering Department.

Zhang, N. L., Poole, D., 1996. Exploiting causal independence in Bayesian network inference. Journal of Artifiacial Intelligence Research 5, 301.

Curriculum Vitae

PERSONAL

Name:	**Plamen PRODANOV**
Nationality:	Bulgarian
Date of Birth:	02.07.1974
Place of Birth:	Varna, Bulgaria
Civil status:	Single
Address:	Av. du Mont d'Or 60, 1007 Lausanne, Switzerland
E-mail:	plamen.prodanov@epfl.ch

EDUCATION

(2002 till now):	Ph.D. in the domain of speech-based human-robot interaction, Autonomous System Lab and LIDIAP, Swiss Federal Institute of Technology - Lausanne (EPFL), Switzerland
(2002):	Postgraduate Course in Computer Science: Language and Speech Engineering, Department of Computer Science, EPFL, Switzerland
(2000-2001):	Doctoral School in Communication Systems, Department of Communication Systems, EPFL, Switzerland
(1993-1998):	Master in Communication and Security Technology and Systems, specializing in Telecommunication Networks, qualification: Electronic and Communication Engineer, Technical University, Varna, Bulgaria
(1988-1993):	High School of Machine Engineering and Electronics, Varna, Bulgaria, majoring in Electronic Technology, special: Industrial Electronics, qualification: Technician

AWARDS

(2005):	Paper [5] nominated among the 10 best papers at the IEEE, EUSIPCO05 conference
(2005):	Best Student Paper award for paper [7] at IEEE, ICASSP05 conference
(2002):	Paper [12] nominated among the 10 best papers (out of 700) at the IEEE, IROS02 conference
(2000-2001) :	Doctoral School Fellowship
(1993-1998):	Fellowship for excellent academic results during all five years of study at the Technical University, Varna

Professional Experience

(2002 till now):	Research and teaching assistantship in the LIDIAP and the Autonomous System Lab, Swiss Federal Institute of Technology - Lausanne, Switzerland. *Responsibilities:* (1). Research and development of multimodal voice-enabled interfaces for mobile service robots; (2). Software development (C++) of systems for multi-sensor data acquisition and processing (including audio, video, laser scanner and buttons input data); (3). Development of multimodal databases; (4). Supervision of laboratory exercises and semester projects.
(04/2000-09/2000):	Software developer (PC and DSP programming (C++, Assembler), software design for embedded radar systems), Laboratory for Signal Processing and Software Systems, Cherno More Co. Varna, Bulgaria.
(07/1999-03/2000):	Software specialist (database development and management), Military service - recruit center, Varna, Bulgaria.
(09/1998-07/1999):	Software developer (PC and DSP programming (C++, Assembler), software design for embedded radar systems), Laboratory for Signal Processing and Software Systems, Cherno More Co. Varna, Bulgaria.
(07/1996-10/1996):	Computer designer and System administrator, GID Consult, Varna, Bulgaria.

Computer skills

Programming:	C/ C++, Visual C++, Visual Basic, Perl, Java, SQL, Pascal, Assembler
Operating Systems:	Windows, Linux, Solaris
Tools:	Matlab, Latex, BNT, HTK, CorelDraw

Language skills

Native:	Bulgarian
Fluent:	English, Russian
Good:	French

Projects

- ♢ Development of voice interface for the mobile tour-guide robots at the Swiss National exhibition (Expo.02), Neuchatel, Switzerland
- ♢ Embedded electronic charts system, developed as part of a radar development project for Sofia Airport, 2000, Cherno More Co. Varna

PUBLICATIONS

1. Drygajlo, A., Prodanov, P., Ramel, G., Meisser, M. and Siegwart, R. (2003). "On Developing a Voice Enabled Interface for Interactive Tour-Guide Robots." In *Advanced Robotics, Robotics Society of Japan*, 2003.

2. Prodanov, Pl., Drygajlo, A., (2005). "Bayesian Networks Based Multimodality Fusion for Error Handling in Human-Robot Dialogues Under Noisy Conditions." *ISCA journal of Speech Communication*, 2005.

3. Richiardi, J., Prodanov, Pl., Drygajlo, A. (2006). "Speaker Verification with Confidence and Reliability Measures." *Proc. ICASSP06*, 2006.

4. Prodanov, Pl., Richiardi, J., Drygajlo, A., (2005). "Graphical Models for Dialogue Repair in Multimodal Interaction with Service Robots." *Proc. of 8th COST 276 Workshop*, May 26-28, 2005, Trondheim, Norway.

5. Kryszczuk, K., Richiardi, J., Prodanov, Pl., Drygajlo, A., (2005). "Error Handling in Multimodal Biometric Systems using Reliability Measures." *in Proc. EUSIPCO05*, 2005.

6. Prodanov, Pl., Drygajlo, A., (2005). "Decision Networks for Repair Strategies in Speech-Based Interaction with Mobile Tour-Guide Robots." *Proc. of Int. Conf. on Robotics and Automation, IEEE ICRA05*, 2005.

7. Richiardi, J., Prodanov, Pl., Drygajlo, A., (2005). "A Probabilistic Measure of Modality Reliability in Speaker Verification." *Proc. IEEE ICASSP05*, 2005.

8. Prodanov, Pl., Drygajlo, A., (2004). "Bayesian Networks for Error Handling through Multimodality Fusion in Spoken Dialogues with Mobile Robots." *Proc. of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 3 October, Jeju, Korea, SAPA.04, 2004.

9. Prodanov, Pl., Drygajlo, A., (2004). "Bayesian Networks Based Signal Fusion for User Goal Identification in Human-Robot Dialogues." *Proc of 6th COST 276 Workshop*, May 6-7, 2004, Thessaloniki, Greece, 2004.

10. Prodanov, P. and Drygajlo, A., (2003). "Multimodal Interaction Management for Tour-Guide Robots Using Bayesian Networks." *Proc. of Int. Conf. on Intelligent Robots and Systems, IEEE IROS'2003*, Las Vegas, USA, 2003.

11. Prodanov, Pl. and Drygajlo, A., (2003). "Bayesian networks for spoken dialogue management in multimodal systems of tour-guide robots." *Proc. of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, Eurospeech 2003.

12. Prodanov, Pl., Drygajlo, A., Ramel, G., Meisser, M., Siegwart, R., (2002)."Voice Enabled Interface for Interactive Tour-Guide Robots." *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems IROS'03*, Lausanne, Switzerland, Sept. 30 - Oct. 4, 2002.